REVOLUSI KESEHATAN DENGAN DATA SCIENCE: MAKSIMALKAN TEKNOLOGI INFORMASI DENGAN RAPIDMINER UNTUK TRANSFORMASI DATA KESEHATAN

Wahyu Wijaya Widiyanto, M.Kom.
Bajeng Nurul Widyaningrum, M.Kom.

Revolusi Kesehatan dengan Data Science: Maksimalkan Teknologi Informasi dengan RapidMiner untuk Transformasi Data Kesehatan

Wahyu Wijaya Widiyanto, M.Kom.
Bajeng Nurul Widyaningrum, M.Kom.

GETPRESS INDONESIA

Revolusi Kesehatan dengan Data Science: Maksimalkan Teknologi Informasi dengan RapidMiner untuk Transformasi Data Kesehatan

Penulis: Wahyu Wijaya Widiyanto, M.Kom.
Bajeng Nurul Widyaningrum, M.Kom.

ISBN: 978-623-198-832-4
Editor: Dina Ediana, M.Kom.
Penyunting: Yuliatri Novita, M.Hum.
Desain Sampul dan Tata Letak: Tri Putri Wahyuni, S.Pd.
Penerbit : GETPRESS INDONESIA
Anggota IKAPI No. 033/SBA/2022

Redaksi :
Jl. Palarik RT 01 RW 06 Kelurahan Air Pacah
Kecamatan Koto Tangah Padang Sumatera Barat

website: www.getpress.co.id
email: adm.getpress@gmail.com

Cetakan pertama, November 2023

KATA PENGANTAR

Selamat datang dalam perjalanan ilmiah yang menggairahkan ke dunia Data Science dalam konteks kesehatan. Buku ini, berjudul " Revolusi Kesehatan dengan Data Science: Maksimalkan Teknologi Informasi dengan RapidMiner untuk Transformasi Data Kesehatan," telah dirancang untuk membawa Anda ke dalam perangkat ilmu data yang memegang peran penting dalam membentuk masa depan perawatan kesehatan.

Di era di mana data kesehatan menjadi lebih melimpah daripada sebelumnya, kita telah memasuki zaman di mana pengetahuan yang mendalam tentang data kesehatan memiliki potensi besar untuk mengubah cara kita merawat dan memahami kesehatan kita. Teknologi informasi dan alat Data Mining, seperti RapidMiner, menjadi pilar yang memungkinkan kita untuk menggali inti dari data tersebut dan menerjemahkannya menjadi wawasan yang bermanfaat.

Penulis buku ini adalah para ahli yang memiliki pengalaman dan pengetahuan mendalam di bidang Data Science dan kesehatan. Dalam buku ini, mereka akan membimbing Anda melalui konsep-konsep dasar Data Mining, teknik-teknik penting, jenis-jenis algoritma, dan bagaimana mengaplikasikannya dalam dunia kesehatan.

Anda juga akan diajak untuk menjelajahi alat-alat Data Mining seperti Weka dan RapidMiner, serta memahami bagaimana menginstal RapidMiner, memproses data, dan menerapkan algoritma dalam studi kasus nyata.

Kami berharap buku ini akan membuka pintu bagi Anda dalam memahami peran penting Data Science dalam perawatan kesehatan dan memberikan Anda alat yang Anda butuhkan untuk menghadapi tantangan dan peluang di dunia yang semakin terhubung secara data ini.

Bergabunglah dengan kami dalam perjalanan ini untuk menggali potensi Data Science dalam memajukan perawatan kesehatan, dan mari bersama-sama menjadi agen perubahan dalam perbaikan kesehatan masyarakat. Selamat membaca!

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
KONSEP DATA MINING

Istilah Data Mining sebenarnya mulai dikenal sejak tahun 1990, ketika pekerjaan pemanfaatan data menjadi sesuatu yang dianggap penting dalam berbagai bidang, mulai dari bidang akademik, bisnis, hingga bidang medis. Munculnya data mining didasarkan pada jumlah data yang tersimpan dalam basis data semakin besar. Perkembangan yang cepat dalam teknologi pengumpulan dan penyimpanan data telah memudahkan organisasi untuk mengumpulkan sejumlah data berukuran besar sehingga menghasilkan gunung data. Misalnya dalam sebuah universitas ada berapa data mahasiswa yang tersimpan dari tahun ke tahun. Kemudian data di sebuah supermarket, ada berapa transaksi pelanggan yang terjadi dalam sehari dan ada berapa juta data yang tersimpan dalam sebulan. Ekstraksi informasi yang berguna dari basis data tersebut menjadi pekerjaan yang cukup menantang. Seringkali alat dan teknik analisis data tradisional tidak dapat digunakan dalam mengekstrak informasi dari data berukuran besar. Data mining adalah teknologi yang merupakan campuran teknik-teknik analisis data dengan algoritmaalgoritma untuk memproses data berukuran besar. Data mining telah banyak diaplikasikan dalam berbagai bidang, seperti dalam bidang bisnis dan kedokteran. Dalam bidang bisnis, teknik data mining digunakan untuk mendukung cakupan yang luas dari aplikasi-aplikasi bisnis inteligen seperti customer profiling, targeted marketing, workflow

management, store layout dan fraud detection. Teknik data mining dapat digunakan untuk menjawab pertanyaan bisnis yang penting seperti “Siapakah pelanggan yang akan paling banyak mendatangkan keuntungan?” dan ”Seperti apa perkiraan pendapatan perusahaan tahun depan?”. Dalam bidang kedokteran, peneliti dalam bidang biomolekuler dapat menggunakan teknik data mining untuk menganalisis sejumlah besar data genomic yang sekarang ini telah banyak dikumpulkan untuk menjelaskan struktur dan fungsi gen, memprediksi struktur protein, dan lain-lain.

1.1 Definisi dan Konsep Data Mining

Kalau kita membahas tentang Data Mining, tentulah kita harus mengetahui terlebih dahulu definisi dari Data Mining. Secara umum Data Mining terbagi atas 2 (dua) kata berikut :

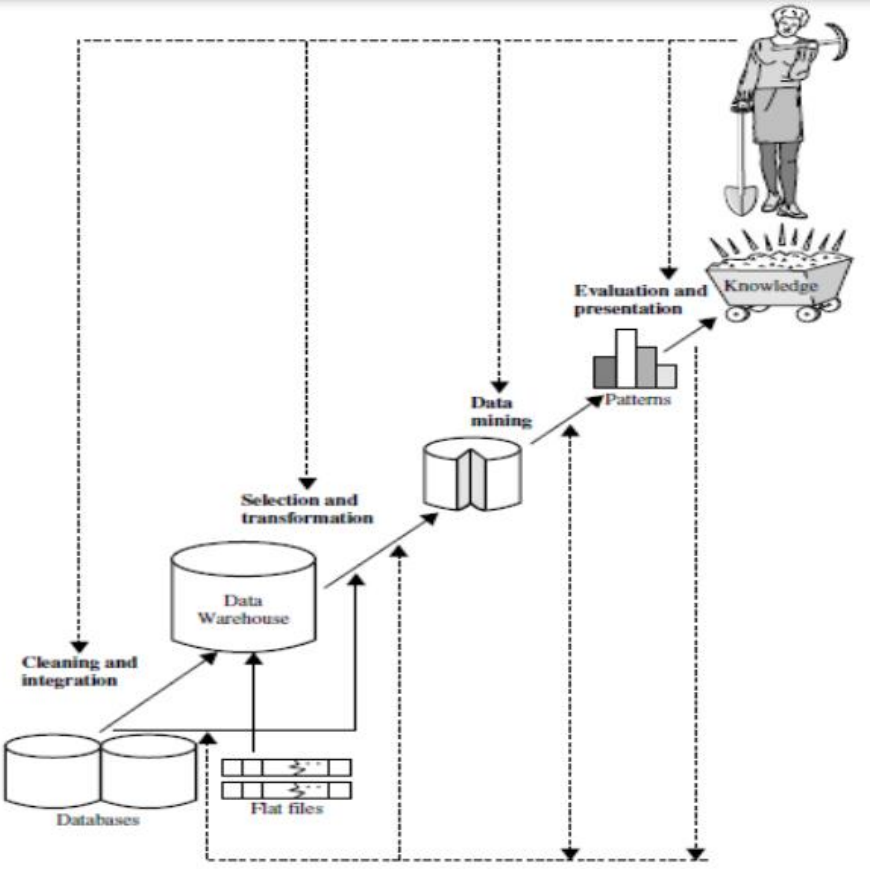
1. Data yaitu Kumpulan Fakta yang terekam atau sebuah entitas yang tidak memiliki arti dan selama ini terabaikan.
2. Mining yaitu proses Penambangan sehingga, Data Mining itu dapat diartikan sebagai proses penambangan data yang menghasilkan sebuah ouput (keluaran) berupa pengetahuan.

Data mining adalah sebuah proses pencarian secara otomatis informasi yang berguna dalam tempat penyimpanan data berukuran besar. Teknik data mining digunakan untuk memeriksa basis data berukuran besar sebagai cara untuk menemukan pola yang baru dan berguna. Namun tidak semua pekerjaan pencarian informasi dapat dinyatakan sebagai data mining. Berikut

ini adalah beberapa contoh yang merupakan data mining dan yang bukan data mining.

1. Bukan Data Mining: Pencarian informasi tertentu di internet, misalnya dengan mengetikkan keyword di mesin pencari. Data Mining: Pengelompokkan informasi yang mirip dalam konteks tertentu pada hasil pencarian dengan mesin pencari. Contoh mengelompokkan hasil pencarian Universitas berdasarkan nilai akreditasi, jumlah mahasiswa, kualitas alumni, dan lain-lain.
2. Bukan Data Mining: Membandingkan data laporan keuangan bulan Januari dan Februari untuk mencari berapa kenaikan permintaan.
   Data Mining: Hasil laporan keuangan saat ini digunakan untuk mempredikasi pengeluaran dan pembelian selanjutnya.

Istilah lain yang sering dikaitkan dengan data mining diantaranya *knowledge discovery (mining) in databases, knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, dan business intelligence*. Data mining adalah bagian integral dari *Knowledge Discovery in Databases (KDD)*. Keseluruhan proses KDD untuk konversi *raw data* (data mentah) ke dalam informasi yang berguna ditunjukkan dalam gambar berikut:



Tahapan Data Mining

Data mining sebenarnya merupakan salah satu bagian proses *Knowledge Discovery in Database (KDD)* yang bertugas untuk mengekstrak pola atau model dari data dengan menggunakan suatu algoritma yang spesifik. Adapun proses KDD sebagai berikut:

1. Data Selection: Pemilihan data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai.
2. Preprocessing: Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning

dengan tujuan untuk membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses enrichment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. Transformation: Proses coding pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam database.
4. Data mining: Proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.
5. Interpretation/Evaluation: Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya atau tidak.

1.2 Data

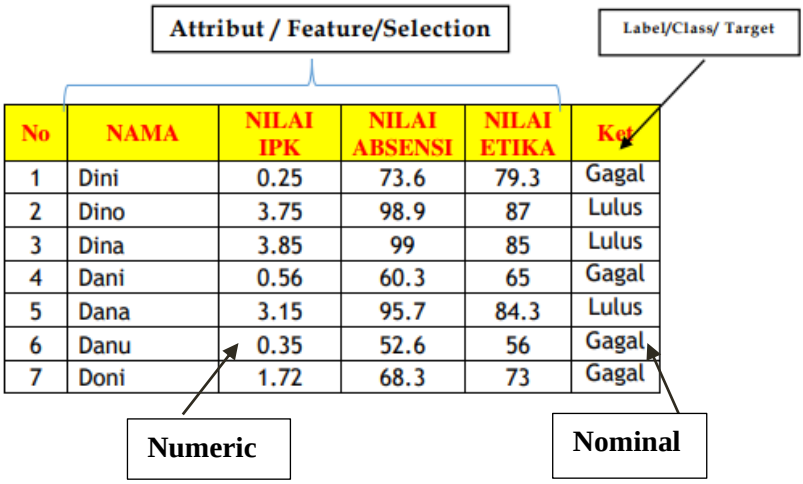
Data mining tidak pernah lepas dari yang namanya data, karena dalam pengolahan data mining, data sangat dibutuhkan sebagai objek untuk mendapakan pengetahuan. Data merupakan kumpulan fakta yang direpresentasikan ke dalam bentuk karakter baik huruf, angka dan lainnya yang dapat diproses menjadi sebuah informasi. Sesuai dengan kaidah penelitian untuk *Data*

Collecting (pengumpulan data) bisa melalui observasi, angket, wawancara dengan stakeholder dan lain-lain. Secara definitif kita mengetahui bahwa Data adalah kumpulan Fakta yang terekam dan tidak memiliki arti. Selain itu data dapat diartikan sebagai kumpulan fakta-fakta yang direpresentasikan kedalam beberapa bentuk baik karakter : Angka, huruf maupun simbol yang diproses sehingga menghasilkan sebuah informasi. Atau data dapat dinterpretasikan sebagai Entitas yang tidak memiliki arti yang selamai ini terabaikan. Data juga dapat di analogi pada dunia pabrikasi yaitu sebagai "Bahan Mentah" sedang hasil pengolahan Produksinya yang disebut "Bahan Jadi" yaitu berupa Informasi. Untuk lebih jelasnya dapat dilihat pada gambar di bawah ini:

Gambar 1.1 Proses Terbentuknya Informasi

Data data mining tentulah kita semua mengetahui bahwa yang akan ditambang atau digali dalam tanda kutip adalah Himpunan Data / Basis Data (database) ,yang kemudian akan diekstraksi menjadi sebuah pengetahuan baik Pola, Klaster, Decision Tree dan lain-lain.Sebelum kita melakukan proses data mining tentunya kita terlebih dahulu mengetahui beberapa elemen dalam sebuah himpunan data seperti pada gambar di bawah ini:

Tabel 1.1 Contoh Himpunan Data



No	NAMA	NILAI IPK	NILAI ABSENSI	NILAI ETIKA	Ket
1	Dini	0.25	73.6	79.3	Gagal
2	Dino	3.75	98.9	87	Lulus
3	Dina	3.85	99	85	Lulus
4	Dani	0.56	60.3	65	Gagal
5	Dana	3.15	95.7	84.3	Lulus
6	Danu	0.35	52.6	56	Gagal
7	Doni	1.72	68.3	73	Gagal

Atribut adalah deskripsi data yang bisa mengidentifikasikan entitas Field adalah lokasi penyimpanan Record adalah kumpulan dari berbagai field yang saling berhubungan.

1. Class / Label / Target bisa disebut sebagai atribut keputusan.
2. Numeric merupakan tipe data yang bisa di kalkulasi
3. Nominal merupakan tipe data yang tidak bisa di kalkulasi baik tambah, kurang, kali maupun bagi.

BAB II
TEKNIK DATA MINING DAN JENIS ALGORITMA

2.1 Teknik Data Mining

Teknik data mining merupakan suatu proses utama yang digunakan saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data. Dengan definisi data mining, ada beberapa teknik dan sifat analisa yang dapat digolongkan dalam data mining yaitu, sebagai berikut:

1. Estimasi

   Teknik untuk melakukan estimasi terhadap sebuah data baru yang tidak memiliki keputusan berdasarkan histori data yang telah ada, dimanavariabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai dari variabel prediksi.
2. Klasifikasi

   Klasifikasi dapat didefinisikan sebagai suatu proses yang melakukan pelatihan/pembelajaran terhadap fungsi target f yang memetakan setiap vektor (set fitur) x ke dalam satu dari sejumlah label kelas y yang tersedia. Di dalam klasifikasi diberikan sejumlah record yang dinamakan training set, yang terdiri dari

beberapa atribut, atribut dapat berupa kontinyu taupun kategoris, salah satu atribut menunjukkan kelas untuk record. Model klasifikasi dapat dilihat pada gambar berikut:

Gambar 2. 1 Blok Diagram Model Klasifikasi

Klasifikasi merupakan teknik yang digunakan untuk menemukan model agar dapat menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Metode klasifikasi yang sering digunakan yaitu, *Support Vector Machine, Multilayer Perceptron, Naive bayes, ID3, Ensemble Methode*, dll.

3. Asosiasi

Teknik untuk mengenali kelakuan dari kejadian-kejadian khusus atau proses dimana hubungan asosiasi muncul pada setiap kejadian. Adapun teknik pemecahan masalah yang sering digunakan seperti Algoritma Apriori. Contoh pemanfaatan Algoritma apriori yaitu pada bidang Marketing ketika sebuah Minimarket melakukan tata letak produk yang dijual berdasarkan produk-produk mana yang paling sering dibeli konsumen,selain itu seperti tata letak buku yang dilakukan pustakawan di perpustakaan. Metode asosiasi yang umum digunakan adalah *FP-Growth, Coefficient of Correlation, Chi Square, A Priori*, dll.

4. Klustering

   Digunakan untuk menganalisis pengelompokkan berbeda terhadap data, mirip dengan klasifikasi, namun pengelompokkan belum didefinisikan sebelum dijalankannya tool data mining. Biasanya menggunkan metode neural network atau statistik, analitikal hierarki cluster. Clustering membagi item menjadi kelompok-kelompok berdasarkan yang ditemukan tool data mining.

5. Prediksi

   Algoritma prediksi biasanya digunakan untuk memperkirakan atau forecasting suatu kejadian sebelum kejadian atau peristiwa tertentu terjadi. Contohnya pada bidang Klimatologi dan Geofisika, yaitu bagaimana Badan Meterologi Dan Geofisika (BMKG) memperkirakan tanggal tertentu bagaimana Cuacanya, apakah Hujan, Panas dan lain sebagainya. Ada beberapa metode yang sering igunakan salah satunya adalah Metode Rough Set. Di dalam data mining juga sama halnya dengan konsep Neural Network mengandung 2(dua) pengelompokkan yaitu:

   a. *Supervised Learning*

      Supervised Learning yaitu pembelajaran menggunakan guru dan biasanya ditandai dengan adanya Class/Label/Target pada himpunan data. Adapun metode-metode yang digunakan yang bersifat supervised learning seperti Metode Prediksi dan Klasifikasi seperti *Decision tree, Nearest - Neighbor Classifier, Naive Bayes Classifier, Artificial Neural Network, Support Vector Machine, Fuzzy K-Nearest Neighbor.*

b. *Unsupervised Learning*

Unsupervised Learning yaitu pembelajaran tanpa *menggunakan* guru dan biasanya ditandai pada himpunan datanya tidak memiliki attribut keputusan atau Class/Label/Target. Adapun metode-metode yang bersifat *Unsupervised Learning* yaitu *K-Means, Hierarchical Clustering, DBSCAN, Fuzzy C-Means, Self-Organizing Map.*

2.2 Jenis-jenis Algoritma

Banyak algoritma yang dapat digunakan dalam pengklasifikasian data, contohnya algoritma *naive bayes, nearest neighbour, dan decision tree*.

1. *Naïve bayes*

Naive Bayes merupakan pengklasifikasian dengan metode probabilitas yang ditemukan oleh ilmuwan Inggris *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Teorema tersebut dikombinasikan dengan naive di mana diasumsikan kondisi antar petunjuk (atribut) saling bebas. Klasifikasi *Naive Bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. Salah satu pengaplikasian dari *Naive Bayes* yaitu pada bidang kesehatan. *Teorema Bayes* adalah perhitungan statistik dengan menghitung probabilitas kemiripan kasus lama yang ada dibasis kasus dengan kasus baru.

Teorema bayes memiliki tingkat akurasi yang tinggi dan kecepatan yang baik ketika diterapkan pada database yang besar. *Naïve bayes* termasuk ke

dalam pembelajaran supervised, sehingga pada tahapan pembelajaran dibutuhkan data awal berupa data pelatihan untuk dapat mengambil keputusan. Pada tahapan pengklasifikasian akan dihitung nilai probabilitas dari masing-masing label kelas yang ada terhadap masukan yang diberikan. Label kelas yang memiliki nilai probabilitas paling besar yang akan dijadikan label kelas data masukan tersebut. *Naive bayes* merupakan perhitungan teorema bayes yang paling sederhana, karena mampu mengurangi kompleksitas komputasi menjadi multiplikasi sederhana dari probabilitas. Selain itu, algoritma *naive bayes* juga mampu menangani set data yang memiliki banyak atribut. Persamaan dari *naive bayes* sebagai berikut:

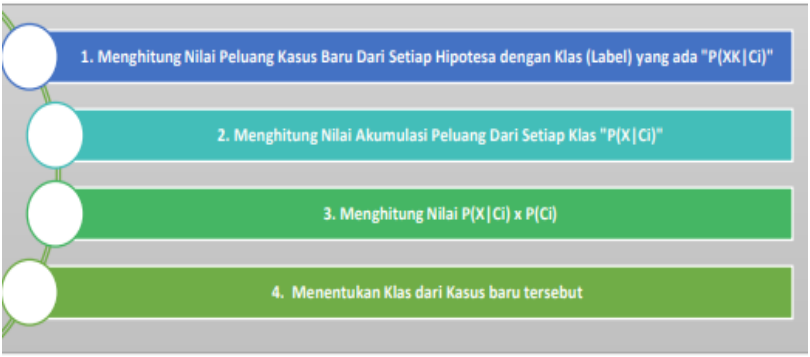
$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad (1)$$

Keterangan :
X : Kriteria suatu kasus berdasarkan masukan
Ci : Kelas solusi pola ke-i, dimana i adalah jumlah label kelas
P(Ci|X) : Probabilitas kemunculan label kelas Ci dengan kriteria masukan X
P(X|Ci) : Probabilitas kriteria masukan X dengan label kelas Ci
P(Ci) : Probabilitas label kelas Ci

Gambar 2. 2 Rumus *Naïve Bayes*

Model *Naïve Bayes* adalah klasifikasi statistik yang dapat digunakan untuk memprediksi suatu kelas. Model *Naïve Bayes* dapat diasumsikan bahwa efek dari suatu nilai atribut sebuah kelas yang diberikan adalah bebas dari atribut-atribut lain.

Naive Bayes memiliki alur seperti pada gambar dibawah ini:



Gambar 2. 3 Alur *Naïve Bayes*

Kelebihan yang dimiliki oleh *Naïve Bayes* adalah dapat menangani data kuantitatif dan data diskrit, *Naïve Bayes* kokoh terhadap noise, *Naïve Bayes* hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter yang dibutuhkan untuk klasifikasi, *Naïve Bayes* dapat menangani nilai yang hilang dengan mengabaikan instansi selama perhitungan estimasi peluang, *Naïve Bayes* cepat dan efisiensi ruang.

Metode ini penting karena beberapa alasan, termasuk berikut. Hal ini sangat mudah untuk membangun, tidak perlu ada yang rumit Parameter estimasi skema berulang. Ini berarti dapat segera diterapkan untuk dataset yang besar. Sangat mudah untuk menafsirkan, sehingga pengguna tidak terampil dalam teknologi classifier dapat memahami.

2. *K-Nearest Neighbor* (KNN)

Nearest Neighbour adalah algoritma pengklasifikasian yang didasarkan pada analogi, yaitu membandingkan data uji dengan data pelatihan

yang berada dekat dengan dan memiliki kemiripan dengan data uji tersebut. KNN merupakan metode yang menggunakan algoritma supervised dimana hasil dari query instance yang baru diklasifikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari algoritma ini adalah mengklasifikasikan obyek baru bedasarkan atribut dan training sample.

Algoritma KNN sangatlah sederhana, bekerja berdasarkan jarak terpendek dari query instance ke training sample untuk menentukan KNN-nya dan mudah untuk di implementasikan. KNN memiliki kemampuan kerja yang rendah ketika training dataset besar. Salah satu masalah pada algoritma ini adalah bobot yang sama dari semua atribut dalam menghitung jarak antara data testing dan data training, bagaimana pun, mungkin dari semua atribut ada beberapa atribut yang kurang penting untuk proses klasifikasi dan ada beberapa atribut yang lebih penting untuk proses klasifikasi. Sehingga tidak jelas jarak mana yang harus digunakan dan atribut mana yang harus digunakan untuk mendapatkan hasil terbaik. Hal ini dapat menyesatkan proses klasifikasi dan dapat menurunkan akurasi dari klasifikasi.

Pendekatan yang banyak dilakukan untuk mengatasi masalah ini adalah dengan memberi bobot yang berbeda pada tiap-tiap atribut ketika mengukur jarak dua record. Pembobotan berguna untuk menentukan jarak antar atribut tetangga dengan record baru berdasarkan similarity. Ketepatan algoritma k-NN ini sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika

bobot fitur tersebut tidak setara dengan relevan isinya terhadap klasifikasi. Riset terhadap algoritma ini sebagian besar membahas bagaimana memilih dan memberi bobot terhadap fitur, agar performa klasifikasi menjadi lebih baik. KNN juga merupakan contoh teknik lazy learning, yaitu teknik yang menunggu sampai pertanyaan (query) datang agar sama dengan data training.

Kemiripan data uji dengan data pelatihan didasarkan pada jaraknya. Banyak persamaan yang dapat digunakan untuk menghitung jarak antara data uji dan data pelatihan. Tiga diantaranya yang paling sering digunakan adalah:

a. Atribut yang bertipe numerik

Terdapat dua pendekatan perhitungan jarak/kemiripan yang umum digunakan untuk atribut yang bertipe numerik, yaitu euclidean distance dengan persamaan berikut:

$$Dist(x1, x2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})} \quad (2)$$

Keterangan:
n : jumlah data
x1 : data uji
x2 : data pembelajaran

Gambar 2. 4 Rumus KNN euclidean distance

Persamaan yang kedua yaitu Manhattan distance sebagai berikut:

$$Dist(p_i(an), p_i(nc)) = \frac{p_i(an) - p_i(nc)}{\max\_dist\_i}$$

Keterangan:
pi : atribut ke-i
an : data pembelajaran
nc : data uji

Gambar 2. 5 Rumus KNN Manhattan distance

b. Atribut yang bertipe simbolik

Persamaan yang digunakan untuk atribut yang menggunakan istilah eksplisit yaitu ada atau tidak ada, memiliki atau tidak memiliki, ya atau tidak dan sebagainya maka perhitungan kemiripan atau jarak dapat dihitung dengan fungsi sebagai berikut:

$$Similarity(T,S) = \frac{\sum_{i=1}^{n} Sim(K_i(T), K_i(S)) x w_i}{\sum_{i=i}^{n} w}$$

Keterangan:
T : Kasus Baru
S : Kasus yang ada dalam penyimpanan
n : jumlah attribut dalam setiap kasus
i : attribut individu antara 1 sampai dengan n
f : fungsi similarity atribut i antara kasus T dan Kasus S
w : bobot yang diberikan pada atribut ke-i

3. *Decision Tree*

Algoritma *decision tree* merupakan algoritma yang umum digunakan untuk pengambilan keputusan. *Decision tree* akan mencari solusi permasalahan dengan menjadikan kriteria sebagai node yang saling berhubungan membentuk seperti struktur pohon. *Decision tree* adalah model prediksi terhadap suatu keputusan menggunakan struktur hirarki atau pohon. Setiap pohon memiliki cabang, cabang mewakili suatu atribut

yang harus dipenuhi untuk menuju cabang selanjutnya hingga berakhir di daun (tidak ada cabang lagi). Konsep data dalam decision tree adalah data dinyatakan dalam bentuk tabel yang terdiri dari atribut dan record. Atribut digunakan sebagai parameter yang dibuat sebagai kriteria dalam pembuatan pohon. Proses dalam decision tree adalah sebagai berikut:

a. Mengubah bentuk data (tabel) menjadi model pohon

   Hal yang dilakukan pada tahapan ini adalah menentukan atribut yang terpilih mulai dari akar, cabang hingga menuju keputusan. Banyak pendekatan yang dapat digunakan untuk menentukan atribut terpilih, pada penelitian ini akan menggunakan perhitungan gainratio dari setiap kriteria dengan data sampel. Untuk menghitung nilai gainratio dapat dilakukan dengan persamaan sebagai berikut:

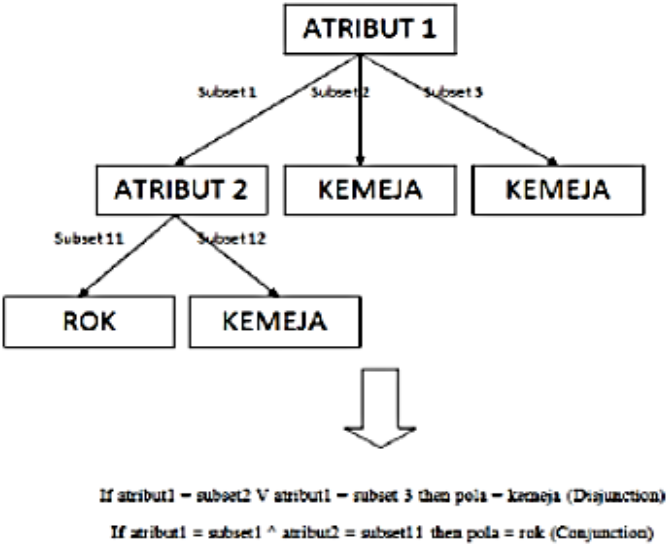
$$Gainratio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)}$$

Gambar 2. 6 Rumus Decision Tree

Dimana nilai information gain bermakna seberapa banyak informasi yang diperoleh dengan mengetahui nilai suatu atribut sedangkan nilai split information digunakan untuk suatu atribut yang memiliki banyak instance (lebih dari dua dan beragam)

b. Mengubah model pohon menjadi rule

Formula untuk membangkitkan rule didefinisikan sebagai berikut: IF premis THEN konklusi Simpul akar dan cabang akan menjadi premis dari aturan, sedangkan simpul daun akan menjadi bagian dari konklusinya (solusi). Tiap premis yang terdapat dalam satu atribut akan dihubungkan dengan hubungan disjungsi, sedangkan premis yang memiliki lanjutan premis pada cabang selanjutnya akan dihubungkan dengan konjungsi.



Gambar 2. 7 Proses Model Pohon Menjadi Rule

c. Menyederhanakan rule (*Pruning*)

Pada proses penyederhanaan rule, tahapan-tahapan dilakukan sebagai berikut:

1) Membuat tabel distibusi terpadu dengan menyatakan semua nilai kejadian pada setiap rule.
2) Menghitung tingkat independensi antara kriteria pada suatu rule, yaitu antara atribut dengan target atribut (perhitungan tingkat independensi menggunakan test of independency Chi-Square).
3) Mengeliminasi kriteria yang dianggap tidak perlu, yaitu yang memiliki tingkat independensi tinggi.

Misalkan yang ingin dilihat adalah pengaruh jenis pakaian terhadap penentuan solusi pola pakaian yang dapat dibuat, tentukan terlebih dahulu tingkat signifikansinya, sehingga dapat dihitung degree of freedom dengan persamaan berikut:

{0.05;(r-1)*(c-1)}

Keterangan:

r : jumlah baris
c : jumlah kolom
Setelah diperoleh nilainya maka dapat dilihat pada tabel untuk memperoleh nilai X^2_{tabel} untuk dibandingkan dengan X^2_{hitung}. X^2_{hitung} diperoleh melalui persamaan berikut :

$$X^2_{hitung} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad (9)$$

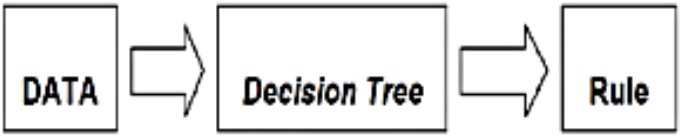
Keterangan:
n_{ij} : nilai *record* baris ke i kolom ke j dari tabel distribusi terpadu.
Sedangkan nilai e_{ij} diperoleh melalui persamaan berikut:

$$e_{ij} = \frac{n_i \bullet n_j}{n} \quad (10)$$

Keterangan:
n_i : marjinal dari baris ke i
n_j : marjinal dari kolom ke j
n : jumlah *record* data

Jika nilai $X^2_{hitung} \leq X^2_{tabel}$ artinya atribut tersebut tidak mempengaruhi atribut target, sehingga *rule* dari atribut tersebut dapat dihilangkan. Namun sebaliknya jika nilai $X^2_{hitung} > X^2_{tabel}$ berarti atribut tersebut mempengaruhi atribut target, sehingga *rule* dari atribut tersebut tidak dapat dihilangkan.

Decision tree adalah bentuk sederhana teknik klasifikasi pada sekumpulan kelas tak berhingga yang direpresentasikan ke dalam bentuk simpul (node) dan rusuk (edge). Biasanya, *Decision Tree* dipilih untuk menyelesaikan masalah dengan output yang bernilai diskrit.



Gambar 2. 8 Konsep Decision Tree

Ada beberapa konsep dalam decision tree, antara lain:

1) Data dinyatakan dalam bentuk Tabel dengan atribut dan record.
2) Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan tree. Misalkan untuk menentukan main tenis, kriteria yang

diperhatikan adalah cuaca, angin dan temperatur. Salah satu atribut merupakan atribut yang menyatakan data solusi per-item data yang disebut dengan target atribut.

3) Atribut memiliki nilai-nilai yang dinamakan dengan instance.

BAB III
KEBUTUHAN AKAN DATA MINING

3.1 Kebutuhan akan Data Mining

Kebutuhan atau pentingnya bidang data mining tidak terlepas dari tujuan dan peran dari data mining itu sendiri. Adapun tujuan data mining secara umum dapat dibagi ke dalam dua kategori utama, yaitu:

1. Prediktif. Tujuannya untuk memprediksi atau memperkirakan nilai dari atribut tertentu berdasarkan pada nilai dari atribut-atribut lain. Atribut yang diprediksi umumnya dikenalsebagai target atau variabel tak bebas,sedangkan atribut-atribut yang digunakan untuk membuat prediksi dikenal sebagai explanatory atau variabel bebas.
2. Deskriptif. Tujuannya untuk menurunkan pola-pola (korelasi, trend, cluster, trayektori, dan anomali) yang meringkas hubungan yang pokok dalam data. Tugas data mining deskriptifsering merupakan penyelidikan dan seringkali memerlukan teknik postprocessing untuk validasi dan penjelasan hasil.

3.2 Penerapan Data Mining

Berikut ini adalah beberapa contoh penerapan teknik data mining dalam berbagai bidang:

1. Kesehatan

   Data mining memiliki potensi besar untuk memperbaiki sistem kesehatan. Menggunakan data

dan analisis untuk mengidentifikasi praktik terbaik yang meningkatkan perawatan dan mengurangi biaya. Peneliti menggunakan pendekatan data mining seperti database multi dimensi, pembelajaran mesin, soft computing, visualisasi data dan statistik. Penambangan dapat digunakan untuk memprediksi volume pasien dalam setiap kategori. Proses dikembangkan untuk memastikan bahwa pasien mendapat perawatan yang tepat, di tempat yang tepat, dan pada saat yang tepat. Data mining juga dapat membantu perusahaan asuransi kesehatan untuk mendeteksi kecurangan dan penyalahgunaan.

2. Analisis Pasar

   Analisis Pasar adalah teknik pemodelan berdasarkan teori bahwa jika seorang membeli kelompok item tertentu, maka cenderung membeli kelompok item lainnya. Teknik ini memungkinkan pengecer memahami perilaku pembelian dari para pembeli. Informasi ini dapat membantu pengecer mengetahui kebutuhan pembeli dan akan mengubah tata letak toko sesuai dengan hasil analisis tersebut. Perbandingan hasil antara toko yang berbeda, atau antara pelanggan dalam kelompok demografis yang berbeda dapat dilakukan dengan menggunakan analisis diferensial.

3. Pendidikan

   Pada bidang pendidikan, muncul bidang baru yang disebut Educational Data Mining (EDM). Bidang tersebut berkaitan dengan teknik pengembangan yang menemukan pengetahuan dari data yang berasal dari lingkungan pendidikan. Tujuan EDM ebagai prediksi perilaku belajar di masa depan siswa,

mempelajari dampak dukungan pendidikan, dan memajukan pengetahuan ilmiah tentang pembelajaran. Data mining dapat digunakan oleh sebuah institusi untuk mengambil keputusan yang akurat dan juga untuk memprediksi hasil siswa. Dengan hasilnya institusi bisa fokus pada apa yang harus diajarkan dan bagaimana cara mengajarnya. Pola belajar siswa dapat diambil dan digunakan untuk mengembangkan teknik mengajar mereka.

4. Rekayasa Manufaktur

   Pengetahuan adalah aset terbaik yang dimiliki perusahaan manufaktur. Alat data mining bisa sangat berguna untuk menemukan pola dalam proses manufaktur yang kompleks. Data mining dapat digunakan dalam perancangan tingkat sistem untuk mengekstrak hubungan antara arsitektur produk, portofolio produk, dan data kebutuhan pelanggan. Ini juga bisa digunakan untuk memprediksi perkembangan produk span time, cost, dan dependencies antar tugas lainnya.

5. CRM

   Customer Relationship Management adalah tentang mengakuisisi dan mempertahankan pelanggan, juga meningkatkan loyalitas pelanggan dan menerapkan strategi yang berfokus pada pelanggan. Untuk menjaga hubungan yang benar dengan pelanggan bisnis, perlu untuk mengumpulkan data dan menganalisa informasi. Di sinilah data mining berperan, dimana dengan teknologi data mining, maka data yang terkumpul dapat digunakan untuk melakukan analisis. Hasil analisis dapat digunakan untuk pengelompokkan

target pemasaran, prediksi pelanggan, profilling pelanggan, dan lainnya.

6. *Fraud Detection* (Deteksi Penipuan)

   Miliaran dolar bisa hilang akibat aksi penipuan. Metode tradisional untuk mendeteksi kecurangan memakan waktu dan kompleks. Data mining membantu dalam memberikan pola yang berarti dan mengubah data menjadi informasi. Setiap informasi yang valid dan berguna adalah pengetahuan. Sistem deteksi kecurangan yang sempurna harus melindungi informasi semua pengguna. Metode yang diawasi mencakup pengumpulan catatan sampel. Catatan ini akan digolongkan sebagai catatan curang atau tidak palsu. Sebuah model dibangun dengan menggunakan data ini dan algoritma dibuat untuk mengidentifikasi apakah rekaman tersebut salah atau tidak.

7. *Intrusion Detection*

   Setiap tindakan yang akan membahayakan integritas dan kerahasiaan sumber daya adalah gangguan. Langkah-langkah defensif untuk menghindari gangguan mencakup otentikasi pengguna, menghindari kesalahan pemrograman, dan perlindungan informasi. Data mining dapat membantu memperbaiki deteksi intrusi dengan menambahkan tingkat fokus pada deteksi anomali. Hal tersebut membantu analis dalam membedakan aktivitas dari aktivitas jaringan biasa sehari-hari. Data mining juga membantu mengekstraksi data yang lebih relevan dengan masalah.

8. Deteksi Kebohongan

   Menangkap penjahat merupakan hal yang mudah, namun membawa keluar kebenaran dari penjahat tersebut adalah hal yang sulit. Penegakan hukum bisa menggunakan teknik penambangan, misalkan untuk menyelidiki kejahatan atau memantau komunikasi tersangka teroris. Hal ini termasuk penambangan teks juga. Proses tersebut berusaha untuk menemukan pola yang berarti dalam data, yang biasanya berupa teks tidak terstruktur. Kemudian sampel data yang dikumpulkan dari penelitian sebelumnya dibandingkan dan dibuat sebuah model untuk deteksi kebekuan.

9. Segmentasi Pelanggan

   Penelitian pasar tradisional dapat membantu untuk meng-segmentasikan pelanggan, namun data mining dapat digunakan untuk meningkatkan efektivitas pasar. Data mining merupakan alat bantu dalam upaya menyelaraskan pelanggan menjadi segmen yang berbeda dan dapat menyesuaikan kebutuhan sesuai pelanggan. Pasar selalu mempertahankan konsumen, sehingga penggunaan data mining memungkinkan untuk menemukan segmen pelanggan berdasarkan kerentanan dan bisnis dapat menawarkannya dengan penawaran khusus dan meningkatkan kepuasan.

10. Perbankan/Keuangan

    Dengan komputerisasi perbankan di mana-mana sejumlah besar data seharusnya dihasilkan dengan transaksi baru. Data mining dapat berkontribusi untuk memecahkan masalah bisnis di bidang perbankan dan keuangan dengan

menemukan pola, sebab-akibat, dan korelasi dalam informasi bisnis dan harga pasar yang tidak segera terlihat oleh manajer karena data volume terlalu besar atau dihasilkan terlalu cepat untuk disaring oleh para ahli. Para manajer dapat menemukan informasi ini untuk segmentasi, penargetan, perolehan, penahanan, dan pemeliharaan pelanggan yang lebih baik.

11. Pengawasan

    Perusahaan Pengawasan perusahaan merupakan pemantauan perilaku seseorang atau kelompok oleh perusahaan. Data yang dikumpulkan paling sering digunakan untuk tujuan pemasaran atau dijual ke perusahaan lain, namun dapat juga dibagi secara reguler dengan instansi pemerintah. Hal ini dapat digunakan oleh bisnis untuk menyesuaikan produk mereka dengan yang diinginkan oleh pelanggan. Data tersebut dapat digunakan untuk tujuan pemasaran langsung, seperti iklan bertarget di Google dan Yahoo, di mana iklan ditargetkan ke pengguna mesin pencari dengan menganalisis riwayat pencarian dan email mereka.

12. Analisis Riset

    Sejarah menunjukkan bahwa kita telah menyaksikan perubahan revolusioner dalam penelitian. Data mining sangat membantu dalam pembersihan data, pra-pengolahan data dan integrasi database. Para peneliti dapat menemukan data serupa dari database yang mungkin membawa perubahan dalam penelitian. Identifikasi sekuens co-occurring dan korelasi antara aktivitas apapun dapat

diketahui. Visualisasi data dan data mining visual memberi kita gambaran yang jelas tentang data.

13. Investigasi Kriminal

    Kriminologi adalah proses yang bertujuan untuk mengidentifikasi karakteristik kejahatan. Sebenarnya analisis kejahatan mencakup menjajaki dan mendeteksi kejahatan dan hubungannya dengan penjahat. Tingginya volume dataset kejahatan dan juga kompleksitas hubungan antar data semacam ini membuat kriminologi menjadi bidang yang tepat untuk menerapkan teknik data mining. Laporan kejahatan berbasis teks dapat diubah menjadi file pengolah kata. Informasi ini bisa digunakan untuk melakukan proses pencocokan kejahatan.

14. Bioinformatika

    Pendekatan Data Mining nampaknya ideal untuk Bioinformatika, karena kaya akan data. Data biologi hasil penambangan membantu untuk mengekstrak pengetahuan yang berguna dari kumpulan data besar yang dikumpulkan dalam biologi, dan bidang ilmu kehidupan lainnya yang terkait seperti kedokteran dan ilmu saraf. Aplikasi data mining untuk bioinformatika meliputi penemuan gen, inferensi fungsi protein, diagnosis penyakit, prognosis penyakit, optimasi pengobatan penyakit, rekonstruksi jaringan interaksi protein dan gen, pembersihan data, dan prediksi lokasi sub-seluler protein.

BAB IV
TOOLS DATA MINING

4.1 *Waikato Environment for Knowledge Analysis* (Weka)

Weka merupakan sebuah tools data mining open source yang berbasis java dan berisi kumpulan algoritma *machine learning* dan data *pre-processing*. *Waikato Environment for Knowledge Analysis* adalah kepanjangan dari Weka yang dibuat di Universitas Waikato, New Zealand yang digunakan untuk pendidikan, penelitian, dan aplikasi eksperimen data mining. Weka mampu menyelesaikan masalah-masalah data mining yang ada di dunia, khususnya klasifikasi yang mendasari pendekatan machine learning. Weka dapat digunakan dan diterapkan pada beberapa tingkatan yang berbeda. Weka menyediakan pengimplementasian algoritma pembelajaran *state of the art* yang dapat diterapkan pada dataset dari command line. Dalam Weka terdapat tools untuk p*reprocessing data, clustering*, aturan asosiasi, *klasifikasi, regres*i, dan visualisasi. Adanya tools yang sudah tersedia pengguna dapat melakukan preprocess pada data, kemudian memasukkannya dalam sebuah skema pembelajaran, dan menganalisis classifier yang dihasilkan serta performanya. Semua itu dilakukan tanpa menulis kode program sama sekali. Contoh penggunaan weka adalah menerapkan sebuah metode pembelajaran ke dataset dan menganalisis hasilnya untuk memperoleh informasi tentang data, atau menerapkan beberapa metode dan membandingkan performa untuk dipilih.

4.2 *Rapid Miner*

Rapid Miner merupakan perangakat lunak yang bersifat terbuka (open source). *Rapid Miner* adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. *Rapid Miner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik.

Rapid Miner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. RapidMiner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi.

Rapid Miner sebelumnya bernama YALE (*Yet Another Learning Environment*), dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh RalfKlinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund. *Rapid Miner* didistribusikan di bawah lisensi AGPL (GNU Affero General Public License) versi 3. Hingga saat ini telah ribuan aplikasi yang dikembangkan mengunakan RapidMiner di lebih dari 40 negara.

Rapid Miner sebagai software open source untuk data mining tidak perlu diragukan lagi karena software ini sudah terkemuka di dunia. *Rapid Miner* menempati peringkat pertama sebagai Software data mining pada polling oleh KDnuggets, sebuah portal data-mining pada 2010-2011. *Rapid Miner* menyediakan GUI (Graphic User Interface) untuk merancang sebuah pipeline analitis. GUI ini akan menghasilkan file XML)Extensible Markup

Language) yang mendefenisikan proses analitis keingginan pengguna untuk diterpkan ke data. File ini kemudian dibaca oleh RapidMiner untuk menjalankan analis secara otomatis. *Rapid Miner* memiliki beberapa sifat sebagai berikut:

1. Ditulis dengan bahasa pemrograman Java sehingga dapat dijalankan di berbagai sistem operasi.
2. Proses penemuan pengetahuan dimodelkan sebagai operator trees
3. Representasi XML internal untuk memastikan format standar pertukaran data.
4. Bahasa scripting memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen.
5. Konsep multi-layer untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
6. Memiliki GUI, command line mode, dan Java API yang dapat dipanggil dari program lain.

Beberapa Fitur dari RapidMiner, antara lain:

1. Banyaknya algoritma data mining, seperti *decision treee* dan *self-organization map*.
2. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, tree chart dan 3D Scatter plots.
3. Banyaknya variasi plugin, seperti text plugin untuk melakukan analisis teks.
4. Menyediakan prosedur data mining dan machine learning termasuk: ETL (*extraction, transformation, loading*), data preprocessing, visualisasi, modelling dan evaluasi

5. Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI
6. Mengintegrasikan proyek data mining Weka dan statistika R.

BAB V
INSTALASI RAPID MINER

5.1 Server System Requirements

Rapid Miner Server menggunakan hard disk untuk menyimpan program itu sendiri, file log, dan data sementara. Mulai dari versi 9 ke atas, *Rapid Miner Server* juga menyimpan data yang diimpor dan model yang dihasilkan di direktori rumah baru yang berada di filesystem instalasi perangkat lunak. Server ini tidak memerlukan lebih dari 1GB ruang hard disk yang tersedia untuk fungsinya sendiri, tergantung pada konfigurasi dan penggunaan log Anda. Namun, direktori rumah dapat menghabiskan ruang yang jauh lebih besar. Oleh karena itu, jumlah ruang disk yang diperlukan tergantung pada jumlah data yang akan disimpan pengguna di repositori server.

Disarankan, jika memungkinkan, untuk mengalokasikan ruang hard disk yang cukup untuk menangani proses dan data yang diunggah oleh pengguna. Ini tidak hanya mendukung penyimpanan informasi yang berhubungan dengan pengguna, tetapi juga membantu dalam pelacakan dan penanganan masalah yang mungkin timbul.

Minimal untuk *Rapid Miner* Server dengan Job Agent harus terpasang:

1. Prosesor dual core dengan kecepatan 2GHz
2. RAM 8GB
3. Ruang hard disk kosong >10GB (sistem berkas harus mendukung UTF-8)
4. IE10 atau yang lebih baru, atau browser yang baik

5. Resolusi browser: 1024x768.

Rekomendasi untuk Server RapidMiner dan setiap Job Agent:

1. Prosesor quad core dengan kecepatan 3GHz atau lebih cepat
2. RAM 32GB hingga 1TB
3. Ruang hard disk yang cukup (sistem berkas harus mendukung UTF-8) agar pengguna dapat menyimpan data di repositori server

Sistem Operasi:

1. Disarankan 64-bit
2. Windows Server 2008 R2, Windows Server 2012, Windows Server 2012 R2, Windows Server 2016, Windows Server 2019
3. Linux

Platform Java:

1. Disarankan 64-bit
2. Oracle Java 8 (Harap dicatat bahwa OpenJDK terbukti tidak kompatibel dalam beberapa konfigurasi)

Database Operasional:

1. PostgreSQL 9.6, 10.8, 11.3
2. Microsoft SQL Server 2017
3. Oracle 12c (tabel terkompresi tidak didukung)
4. MySQL 5.6, 5.7, 8.0.

5.2 Download dan Instalasi *Rapid Miner*

1. Download *Rapid Miner*

Seperti yang yang telah dikemukakan sebelumnya bahwa RapidMiner merupakan software gratis yang bersifat terbuka (open source). Software ini dapat dijalankan pada sistem operai Windows, Linux, maupun Mac. RapidMiner dapat diunduh pada situs resminya, yaitu https://rapidminer.com/platform/ . Pada bagian ini, akan dijelaskan bagaimana cara melakukan instalasi software RapidMiner versi 10.1 pada sistem operasi Microsoft Windows

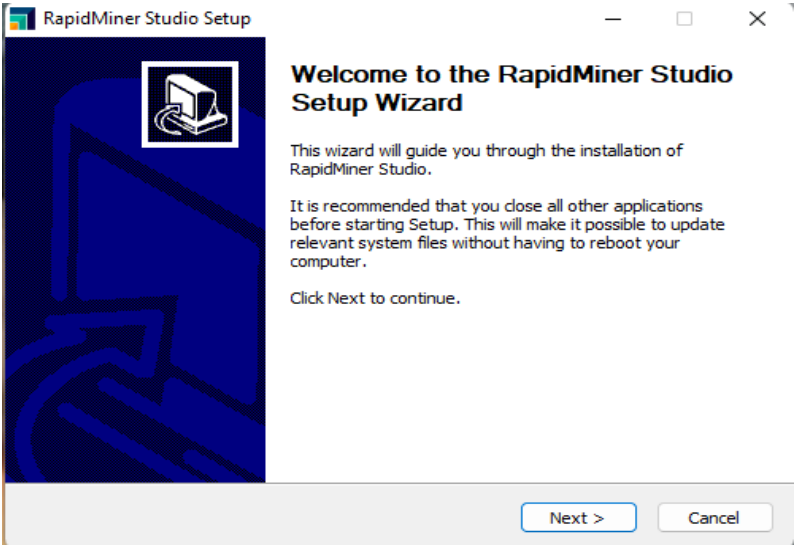
Ikuti petunjuk ini untuk mengunduh RapidMiner Studio:

a. Untuk mengunduh aplikasi, pergi ke situs web RapidMiner.
b. Klik Akun di sudut kanan atas.
c. Masuk jika Anda belum melakukannya. Lihat di bawah jika Anda perlu membuat akun. Jika Anda hanya ingin mengunduh RapidMiner Studio tanpa masuk, klik Unduhan.
d. Klik pada sistem operasi pilihan Anda untuk memulai pengunduhan. Sistem operasi Anda saat ini akan di-highlight.

2. Instalasi *RapidMiner*

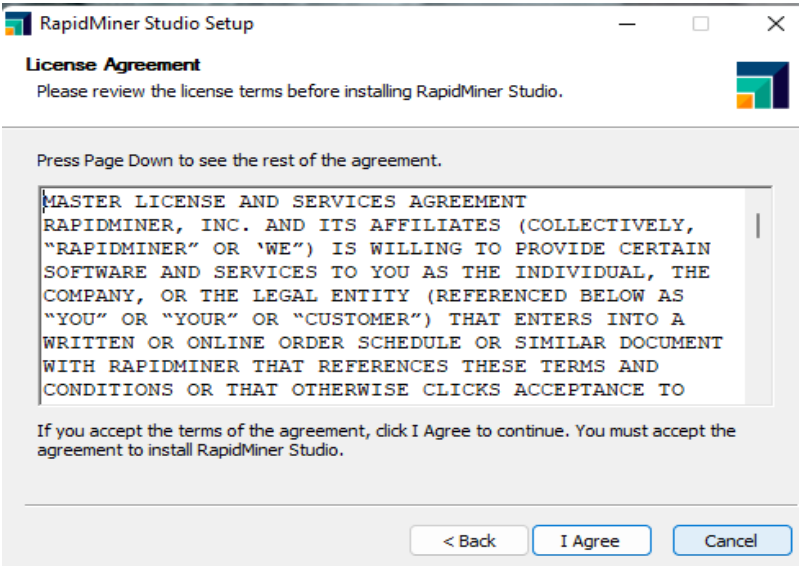
Untuk memulai instalasi software *RapidMiner* pada sistem operasi Microsoft Windows, jalankan file installer rapidminer-studio-10.1.3-win64-install.

a. Klik dua kali file executable dari RapidMiner 10.1.3-win64-install



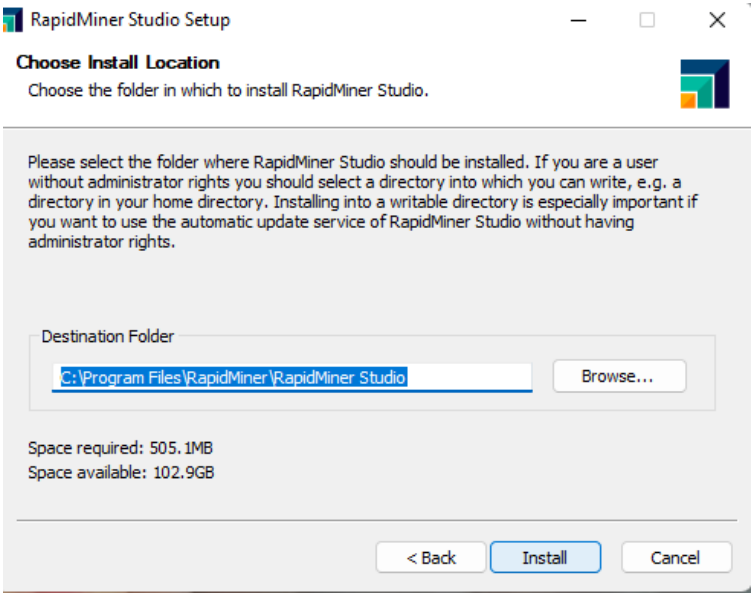
Gambar 5. 1 Form Awal Instalasi

b. Klik Next > untuk melanjutkan pada form persetujuan dan lisensi



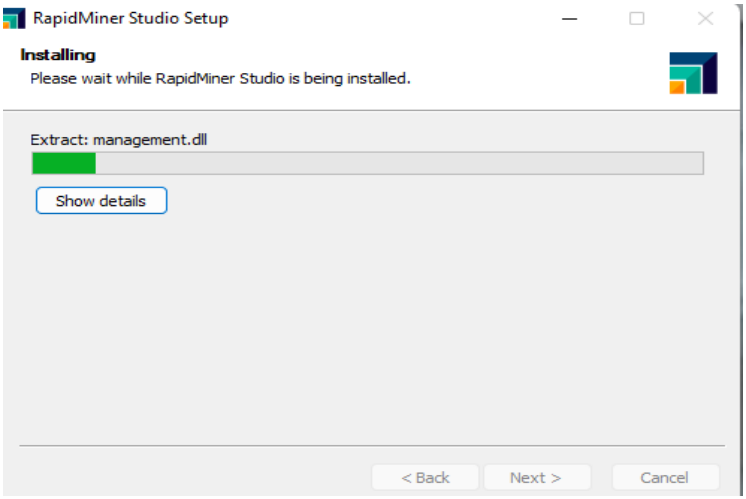
Gambar 5. 2 Form Persetujuan Lisensi

c. Pilih I Agree untuk melanjutkan. Kemudian, wizard akan menampilkan form



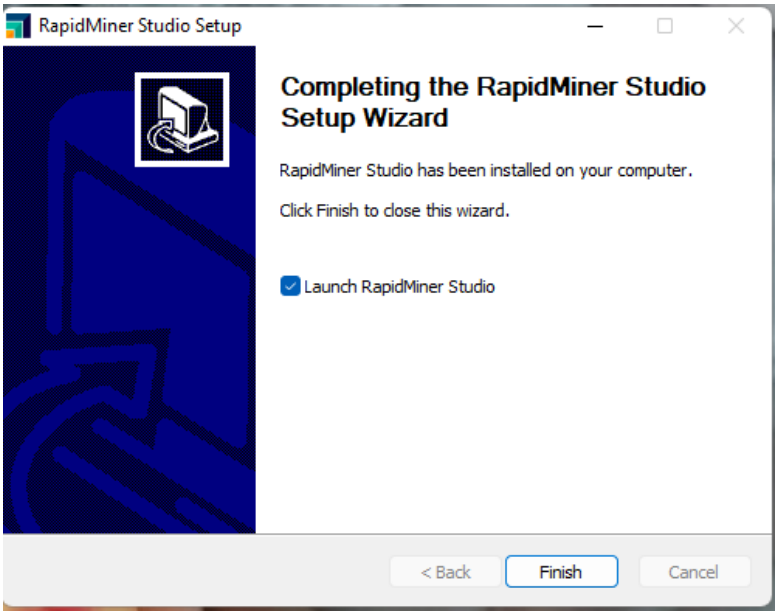
Gambar 5. 3 Form Pemilihan Lokasi Instalasi

Pilih Install untuk melakukan proses instalasi. Kemudian wizard akan menampilkan progress dari proses tersebut



Gambar 5. 4 Form Proses Instalasi

d. Setelah proses selesai, pilih Next > untuk melanjutkan, maka wizard akan menampilkan informasi bahwa proses instalasi telah selesai dilakukan.



Gambar 5. 5 Form Instalasi selesai

Pilih Finish untuk mengakhiri proses instalasi. Lalu akan muncul tampilan awal dari *RapidMiner* versi 10.1

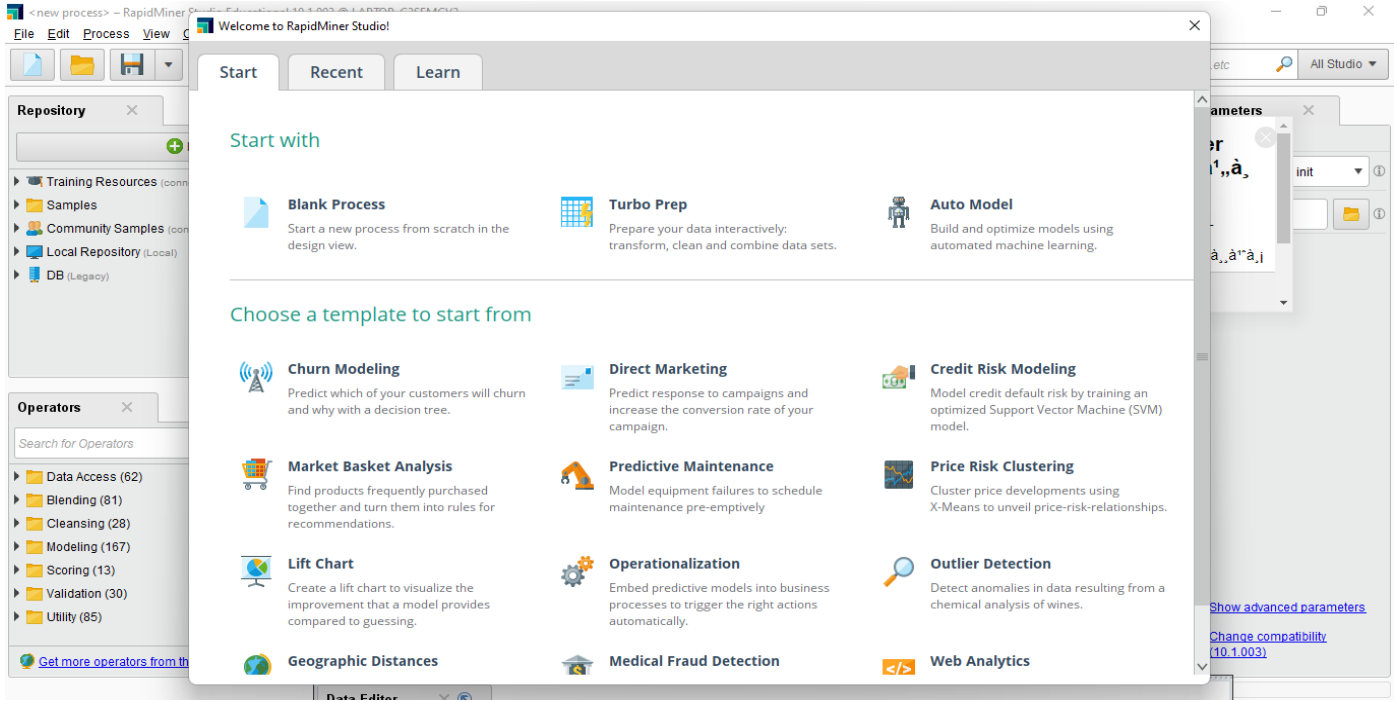
Gambar 5. 6 Tampilan Awal RapidMiner

5.3 Pengenalan Interface *RapidMiner*

RapidMiner menyediakan tampilan yang user friendly untuk memudahkan penggunanya ketika menjalankan aplikasi. Tampilan pada RapidMiner dikenal dengan istilah perspective dan terdapat 3 perspective dalam *RapidMiner* diantaranya *welcome perspective*, design perspective, dan result perspective.

1. ***Welcome Perspective***

   Welcome Perspective Ketika membuka aplikasi akan disambut dengan tampilan yang disebut dengan *Welcome Perspective.* Pada bagian toolbar, terdapat *toolbar Perspectives* yang terdiri dari ikon-ikon untuk menampilkan perspective dari RapidMiner. Toolbar ini dapat dikonfigurasikan sesuai dengan kebutuhan. Terdapat menu start, recent, dan learn. Tampilan perspektif dan tampilan ini ditampilkan saat membuka aplikasi rapidminer untuk pertama kalinya.



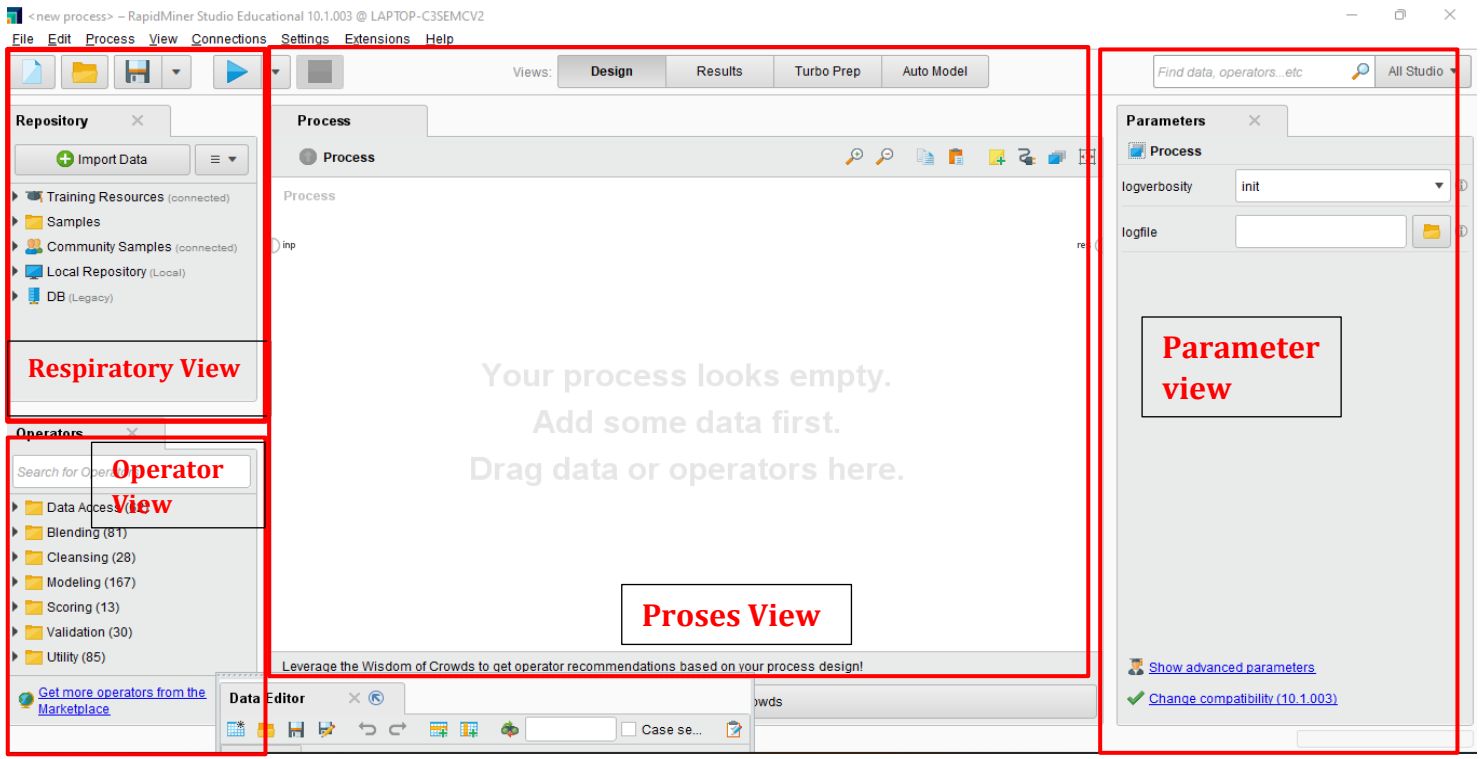
Gambar 5. 7 Welcome Perspective

Pada gambar di atas menampilkan beberapa daftar aksi, diantaranya Start, Recent, dan Learn. Berikut ini rincian lengkap daftar aksi tersebut:

a. Start adalah menu yang digunakan untuk menampilkan beberapa tindakan yang digunakan untuk proses tersebut.
b. Recent, Digunakan untuk menampilkan proses yang baru saja dan sebelumnya dijalankan.
c. Learn digunakan untuk melihat konten pembelajaran yang lebih detail tentang penggunaan Rapidminer yang disediakan oleh aplikasi.

2. Design *Perspective*

Design *perspective* merupakan lembar kerja pada *RapidMiner* yang digunakan untuk membuat dan mengelola proses analisis dari *konsep* data mining



Gambar 5. 8 Tampilan Design Perspertive

Berikut rincian dari beberapa view yang ditampilkan pada design perspective:

a. Operator View

Operator view merupakan salah satu view yang paling penting karena semua operator dari RapidMiner disajikan dalam bentuk hierarki sehingga operator-operator tersebut dapat digunakan pada proses analisis data mining. Di dalam operator view terdapat:

1) Data Access, Untuk menampilkan dan menulis *repository*
2) Blending, berfungsi untuk transformasi data dan meta data
3) Cleansing, Untuk tahapan pra proses data
4) Modelling, Untuk proses data mining, seperti asosiasi, estimasi, klasifikasi, klastering dll.
5) Scoring, untuk pengambilan model dari modelling
6) Validation, Untuk mengutur kualitas dan permormansi dari model atau modelling.
7) Utility, Untuk mengelompokkan subprocess, juga macro dan logger
8) Extensions, untuk format impor dan ekspor dari berbagai data berformat eksternal

b. Process View

Process view merupakan halaman yang digunakan untuk menunjukkan langkah-langkah dalam proses analisis data mining dengan menggunakan komponen-komponen yang ada pada operator view.

c. Parameter View

Beberapa operator dalam RapidMiner membutuhkan satu atau lebih parameter agar dapat didefinisikan sebagai fungsionalitas yang benar. Namun terkadang parameter tidak mutlak dibutuhkan, meskipun eksekusi operator dapat dikendalikan dengan menunjukkan nilai parameter tertentu.

d. Repository View
   Repository view merupakan komponen utama dalam design perspective. Repository view digunakan untuk mengelola dan menata proses analisis data mining.
e. Data Editor
   Data Editor dalam RapidMiner adalah komponen yang memungkinkan Anda untuk memasukkan, mengubah, dan mengelola data yang akan digunakan dalam proses analisis. Ini adalah tempat di mana Anda dapat memanipulasi dataset dengan mengedit, membersihkan, dan mengatur variabel-variabel serta nilai-nilai dalam data tersebut.

3. Result *Perspective*

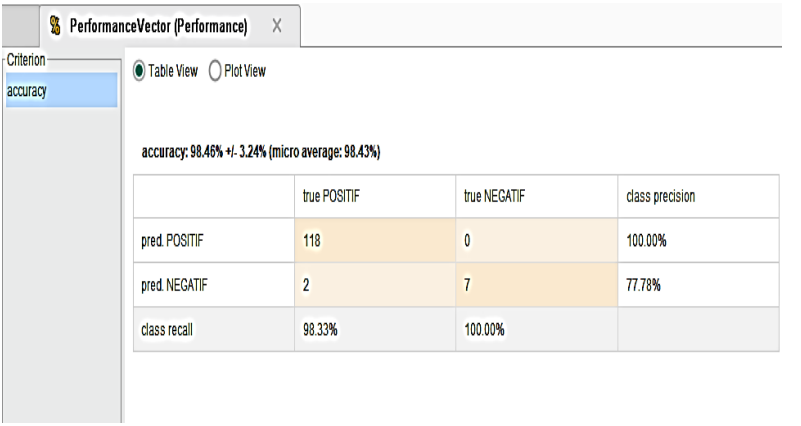
Fitur ini mungkin merujuk pada tampilan atau antarmuka dalam RapidMiner Studio yang didedikasikan untuk menampilkan hasil dari analisis data yang telah dijalankan menggunakan alur kerja sebelumnya. Namun, secara umum, dalam "Result Perspective," Anda mungkin menemukan fitur-fitur seperti:

a. Hasil Analisis: Tampilan visual dan laporan hasil dari eksekusi alur kerja analisis data. Ini

mungkin meliputi grafik, tabel, statistik, dan hasil evaluasi model.

b. Interpretasi Hasil: Penjelasan atau interpretasi tentang apa yang dapat diambil dari hasil analisis. Ini mungkin meliputi ringkasan temuan kunci, insight, dan rekomendasi.
c. Penyusunan Laporan: Alat untuk membuat laporan yang dapat diunduh atau dibagikan dengan pihak lain. Ini mungkin termasuk opsi untuk menyisipkan visualisasi, tabel, dan teks naratif.
d. Interaksi dengan Hasil: Kemungkinan adanya opsi untuk mengeksplorasi dan berinteraksi dengan hasil, seperti memfilter atau menyortir data dalam grafik atau tabel.
e. Visualisasi Lanjutan: Opsi untuk menghasilkan visualisasi yang lebih mendalam atau kustom dari hasil analisis.

   Contoh result perspective ditunjukkan pada gambar di bawah ini:



Gambar 5. 9 Contoh result perspective

5.4 Tutorial Naïve Bayes dengan RapidMiner

Berikut adalah panduan langkah demi langkah untuk membuat analisis klasifikasi Naive Bayes menggunakan RapidMiner:

1. Persiapan Data:
   - Buka RapidMiner Studio.
   - Impor dataset yang akan Anda gunakan untuk analisis. Ini bisa berupa file CSV, Excel, atau sumber data lainnya.
2. Membuat Proses Analisis:
   - Buat proses analisis baru dengan mengeklik ikon "+" di bagian kiri atas jendela.
3. Input Data:
   - Seret operator "Retrieve" dari panel ke bagian kerja.
     - Klik dua kali pada operator "Retrieve" dan pilih dataset yang diimpor sebelumnya.

4. *Preprocessing* Data (Opsional):
    - Anda dapat menambahkan langkah-langkah preprocessing seperti menghapus kolom yang tidak diperlukan, mengisi nilai-nilai yang hilang, atau melakukan transformasi data jika diperlukan.
5. *Naive Bayes* Operator:
    - Seret operator "*Naive Bayes*" dari panel ke bagian kerja.
    - Sambungkan data "Training" dan data “Testing” ke "Naive Bayes".
6. Evaluasi Model:
    - Seret operator "*Performance (*Classification)" ke bagian kerja.
    - Sambungkan output "*Model*" dari operator "Naive Bayes" ke input "Model" dari operator "*Performance* (*Classification*)".
7. Eksekusi dan Analisis Hasil:
    - Klik tombol "Run" di bagian bawah jendela untuk menjalankan proses analisis.
    - Setelah proses selesai, Anda dapat melihat hasil performa model Naive Bayes pada operator "*Performance* (*Classification*)".
8. Interpretasi Hasil:
    - Analisis hasil performa model, termasuk metrik seperti akurasi, presisi, recall, dan F1-score.
    - Jika hasil kurang memuaskan, Anda dapat kembali ke langkah-langkah sebelumnya untuk melakukan tuning atau preprocessing yang lebih baik.
9. Pelaporan Hasil:

- Jika Anda puas dengan hasilnya, Anda dapat menyimpan atau mengekspor model yang telah Anda buat untuk digunakan pada data baru.

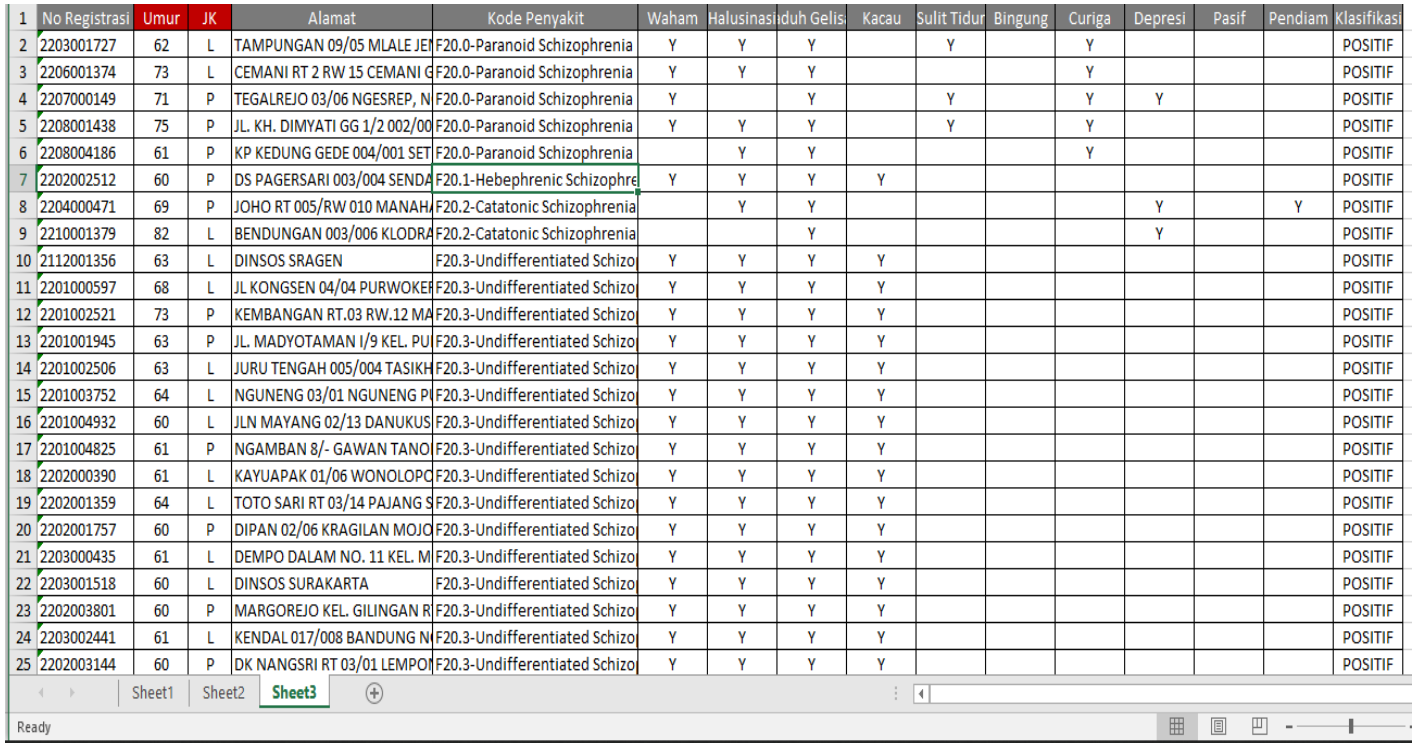
BAB VI
CONTOH KASUS

Pemahaman data

Proses data mining diawali dengan melakukan pemahaman data, yang bertujuan untuk memahami karakteristik data yang akan diolah. Adapun data yang digunakan dalam penelitian ini yaitu data pasien dari RSJD XYZ mengenai pasien yang terdiagnosa penyakit Skizofrenia pada geriatri tahun 2022 sebanyak 127 data. Yang terdiri dari 120 data yang memiliki gejala positif dan 7 data memiliki gejala negatif. Data tersebut diambil dari pasien skizofrenia paranoid, skizofrenia hibefrenik, skizofrenia katatonik, skizofrenia residual, skizofrenia tidak terperinci, skizofrenia residual, dan skizofrenia lainnya. Sumber data yang diambil masih secara manual dengan melihat rekam medis pasien dan diinput menggunakan *Excel* dimana data ini akan di konversi kedalam *RapidMiner*, penjelasan sumber data sebagai berikut:

1. **Data Pasien (Data *Selection*)**

   Data yang berhubungan dengan keluhan pasien mulai dari tanggal registrasi, umur, jenis kelamin, alamat, diagnosis penyakit, dan keluhan pasien.



	No Registrasi	Umur	JK	Alamat	Kode Penyakit	Waham	Halusinasi	duh Gelis	Kacau	Sulit Tidur	Bingung	Curiga	Depresi	Pasif	Pendiam	Klasifikasi
2	2203001727	62	L	TAMPUNGAN 09/05 MLALE JE	F20.0-Paranoid Schizophrenia	Y	Y	Y		Y		Y				POSITIF
3	2206001374	73	L	CEMANI RT 2 RW 15 CEMANI G	F20.0-Paranoid Schizophrenia	Y	Y	Y				Y				POSITIF
4	2207000149	71	P	TEGALREJO 03/06 NGESREP, N	F20.0-Paranoid Schizophrenia	Y		Y		Y		Y	Y			POSITIF
5	2208001438	75	P	JL. KH. DIMYATI GG 1/2 002/00	F20.0-Paranoid Schizophrenia	Y	Y	Y		Y		Y				POSITIF
6	2208004186	61	P	KP KEDUNG GEDE 004/001 SET	F20.0-Paranoid Schizophrenia		Y	Y				Y				POSITIF
7	2202002512	60	P	DS PAGERSARI 003/004 SENDA	F20.1-Hebephrenic Schizophre	Y	Y	Y	Y							POSITIF
8	2204000471	69	P	JOHO RT 005/RW 010 MANAH	F20.2-Catatonic Schizophrenia		Y	Y					Y		Y	POSITIF
9	2210001379	82	L	BENDUNGAN 003/006 KLODRA	F20.2-Catatonic Schizophrenia			Y					Y			POSITIF
10	2112001356	63	L	DINSOS SRAGEN	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
11	2201000597	68	L	JL KONGSEN 04/04 PURWOKE	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
12	2201002521	73	P	KEMBANGAN RT.03 RW.12 MA	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
13	2201001945	63	P	JL. MADYOTAMAN I/9 KEL. PU	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
14	2201002506	63	L	JURU TENGAH 005/004 TASIKH	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
15	2201003752	64	L	NGUNENG 03/01 NGUNENG P	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
16	2201004932	60	L	JLN MAYANG 02/13 DANUKUS	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
17	2201004825	61	P	NGAMBAN 8/- GAWAN TANO	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
18	2202000390	61	L	KAYUAPAK 01/06 WONOLOPO	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
19	2202001359	64	L	TOTO SARI RT 03/14 PAJANG S	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
20	2202001757	60	P	DIPAN 02/06 KRAGILAN MOJO	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
21	2203000435	61	L	DEMPO DALAM NO. 11 KEL. M	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
22	2203001518	60	L	DINSOS SURAKARTA	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
23	2202003801	60	P	MARGOREJO KEL. GILINGAN R	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
24	2203002441	61	L	KENDAL 017/008 BANDUNG N	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF
25	2202003144	60	P	DK NANGSRI RT 03/01 LEMPO	F20.3-Undifferentiated Schizo	Y	Y	Y	Y							POSITIF

Gambar 6. 1 Data Selection

Adapun setiap atribut diuraikan pada tabel di bawah ini:

Tabel 6. 1 Atribut Data Selection

No	Nama Atribut	Keterangan
1.	No Registrasi	Nomor pendaftaran pasien
2.	Umur	Lamanya hidup dalam tahun yang dihitung sejak dilahirkan
3.	Jenis Kelamin	Perbedaan bentuk, sifat dan fungsi biologis antara laki-laki dan perempuan
4.	Alamat	Identitas tempat tingga;
5.	Diagnosa	Penentuan jenis penyakit
6.	Gejala	Keluhan yang dialami

2. Persiapan Data

Jumlah data yang diperoleh pada penelitian ini yaitu 127 data. Data tersebut terdiri dari 120 data yang memiliki gejala positif dan 7 data memiliki gejala negatif.

a. Data *Cleaning*

Untuk mengidentifikasi dan menghapus anomali data atau inkonsisten data (Heliyanti Susana, 2022). Keadaan tersebut bisa terjadi karena dapat *misssing value.* Tahapan data *cleaning* dilakukan sebagai berikut:

JK	ode Penyal	Waham	Halusinasi	Gaduh Gelisah	Kacau	Sulit Tidur	Bingung	Curiga	Depresi	Pasif	Pendiam	Klasifikasi
L	F20.0-Para	Ya	Ya	Ya	Tidak	Ya	Tidak	Ya	Tidak	Tidak	Tidak	POSITIF
L	F20.0-Para	Ya	Ya	Ya	Tidak	Tidak	Tidak	Ya	Tidak	Tidak	Tidak	POSITIF
L	F20.6-Sim	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Ya	Ya	Tidak	NEGATIF
P	F20.6-Sim	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Ya	Ya	Ya	NEGATIF
P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
P	F20.1-Heb	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
P	F20.2-Cata	Tidak	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Ya	Tidak	Ya	POSITIF
L	F20.2-Cata	Tidak	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Ya	Tidak	Tidak	POSITIF
L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
P	F20.5-Resi	Tidak	Ya	Tidak	Tidak	Ya	Tidak	Tidak	Tidak	Ya	ya	NEGATIF
P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF

Gambar 6. 2 Tahapan Data Cleaning

b. Data *Transformation*

Untuk meningkatkan akurasi dan efisiensi algoritma. Data dalam penelitian ini dikategorikan atau diklasifikasi (Loelianto et al., 2020). Untuk data *tranformation* ini misalnya pada bagian kelamin, laki-laki diubah menjadi 0 dan wanita diubah menjadi 1.

Tabel 6. 2 Data Transformation

No	**Atribut**	**Sub Atribut**	**Konversi**
1.	Jenis kelamin	0	Laki-laki
		1	Perempuan
2.	Diagnosa	1	F20.0 (Paranoid Schizophrenia)
		2	F20.1 (Hebephrenic Schizophrenia)
		3	F20.2 (Catatonic Schizoprenia)
		4	F20.3 (Undifferentiated Schizophrenia)
		5	F20.6 (Simple Schizophrenia)
		6	F20.8 (Other Schizophrenia)
3	Waham	0	Ya
		1	Tidak
	Halusinasi	0	Ya
		1	Tidak
	Gaduh gelisah	0	Ya
		1	Tidak
	Kacau	0	Ya
		1	Tidak
	Sulit Tidur	0	Ya
		1	Tidak
	Bingung	0	Ya
		1	Tidak
	Curiga	0	Ya
		1	Tidak

No	Atribut	Sub Atribut	Konversi
	Depresi	0	Ya
		1	Tidak
	Pasif	0	Ya
		1	Tidak
	Pendiam	0	Ya
		1	Tidak
	Gejala	0	Positif
		1	Negatif

Sumber:Data Primer

Setelah proses transformasi selesai maka dataset telah siap untuk diolah ke tahap selanjutnya. Berikut merupakan dataset yang telah melalui transformasi data:

JK	Kode	Waham	Halusinasi	Gaduh Gelisah	Kacau	Sulit Tidur	Bingung	Curiga	Depresi	Pasif	Pendiam	Klasifikasi
0	1	1	1	1	0	1	0	1	0	0	0	1
0	1	1	1	1	0	0	0	1	0	0	0	1
0	6	0	1	0	0	0	0	0	1	1	0	2
1	6	0	0	0	0	0	0	0	1	1	1	2
1	3	1	1	1	1	0	0	0	0	0	0	1
1	1	1	1	1	1	0	0	0	0	0	0	1
1	2	0	1	1	0	0	0	0	1	0	1	1
0	2	0	0	1	0	0	0	0	1	0	0	1
0	3	1	1	1	1	0	0	0	0	0	0	1
0	3	1	1	1	1	0	0	0	0	0	0	1
1	5	0	1	0	0	1	0	0	0	1	1	0
1	3	1	1	1	1	0	0	0	0	0	0	1
0	3	1	1	1	1	0	0	0	0	0	0	1
0	3	1	1	1	1	0	0	0	0	0	0	1
0	3	1	1	1	1	0	0	0	0	0	0	1
1	3	1	1	1	1	0	0	0	0	0	0	1
0	3	1	1	1	1	0	0	0	0	0	0	1
0	3	1	1	1	1	0	0	0	0	0	0	1
1	3	1	1	1	1	0	0	0	0	0	0	1
0	3	1	1	1	1	0	0	0	0	0	0	1
0	3	1	1	1	1	0	0	0	0	0	0	1
1	3	1	1	1	1	0	0	0	0	0	0	1
0	3	1	1	1	1	0	0	0	0	0	0	1
1	3	1	1	1	1	0	0	0	0	0	0	1

Gambar 6. 3 Dataset Telah Melalui Transformasi Data

3. Data *Training* dan Data Testing

Untuk memperoleh data set dengan jumlah atribut yang lebih sedikit tetapi bersifat informatif. Maka data set dibagi menjadi data *training* dan data *testing*. Data training yaitu data yang memiliki atribut label/kelas yang digunakan untuk mengenal karaktersitik data sehingga menghasilkan sebuah pola. Sementara data

testing yaitu data yang juga memiliki label/kelas yang digunakan untuk menguji ketepatan pola/model (Musu et al., 2021). Data *testing* diperoleh dari pembagian sumber data dengan metode 10-*fold cross selection* dimana setiap 10 data latih terdapat 1 data uji. Sehingga pada penelitian ini dengan jumlah dataset sebesar 127, terdapat 115 data latih dan 12 data uji.

4. Pemodelan

Merupakan tahapan implementasi dari teknik dan lagoritma yang dipilih. Dalam penelitian ini dilakukam analisa menggunakan klasifikasi dengan algoritma naïve bayes. *Tools* yang akan digunakan untuk melakukan perhitungan manual menggunakan *Microsoft Excel*, sedangkan untuk pengujiannya menggunakan *software Rapid Miner*. Jumlah sampel pasien penyakit skizofrenia yang dijadikan sampel dalam proses perhitungan algoritma naïve bayes untuk mengetahui pasien tersebut terklasifikasi menderita penyakit skizofrenia bergejala positif atau negatif yang dilihat berdasarkan probabilitas dari setiap kelas dan probabilitas setiap kejadian perkelas dan keakurasian yang dihasilkan.

a. Pemodelan Menggunakan Naïve Bayes

1) Menghitung Probabilitas Setiap Kelas

Dari data yang digunakan terdapat 2 class yaitu positif dan negatif. *Class* Positif berjumlah 109 data, sedangkan *Class*

Negatif berjumlah 6 data. Untuk menghitung probabilitas pada setiap kelas.

Tabel 6. 3 Menghitung Probabilitas Setiap

Kelas	Probabilitas **kelas P(H)**
Positif	109/115=0,9478
Negatif	6/115 = 0,0521

2) Menghitung Probabilitas Setiap Kejadian Perkelas

Untuk menghitung nilai probabilitas setiap atribut dengan cara menghitung jumlah kejadian atau atribut pada suatu kelas dibagi dengan kelas yang ada.

Tabel 6. 4 Perhitungan Probabilitas Setiap Kejadian Perkelas

No	Atribut	Sub Atribut	Probabilitas P(X|H)	
			Positif	Negatif
1.	Jenis kelamin	Laki-laki	65/109=0,5963	3/6 = 0,5
		Perempuan	44/109= 0,4036	3/6 = 0,5
2.	Diagnosa	F20.0(Paranoid Schizophrenia)	5/109= 0,0458	0/6 = 0
		F20.1(Hebephrenic Schizophrenia)	1/109= 0,0091	0/6 = 0
		F20.2(Catatonic Schizoprenia)	2/109= 0,0183	0/6 = 0
		F20.3(Undifferentiated Schizophrenia)	101/109= 0,9266	0/6 = 0
		F20.5(Residual Schizophrenia)	0/109 = 0	2/6 = 0,3333
		F20.6 (Simple Schizophrenia)	0/109 = 0	4/6 = 0,6666
		F20.8 (Other Schizophrenia)	2/109 = 0,0183	0/6 = 0
3	Waham	Ya	106/109= 0,9724	1/6 = 0,1666

		Tidak	1/109 =0,0091	5/6 = 0,8333
	Halusinasi	Ya	108/109=0,9908	2/6 = 0,3333

Tabel 6. 4 Perhitungan Probabilitas Setiap Kejadian Perkelas(Lanjutan)

No	Atribut	Sub Atribut	Probabilitas P(X|H)	
			Positif	Negatif
		Tidak	1/109= 0,0091	3/6 = 0,5
	Gaduh gelisah	Ya	109/109 = 1	0/6 = 0
		Tidak	0/109 = 0	6/6 = 1
	Kacau	Ya	102/109= 0,9357	0/6 = 0
		Tidak	7/109 = 0,642	6/6 = 1
	Sulit Tidur	Ya	5/109 = 0,458	2/6 = 0,3333
		Tidak	104/109= 0,9541	4/6 = 0,6666
	Bingung	Ya	0/109=0	2/6 = 0,3333
		Tidak	107/109=0,9816	6/6 = 1
	Curiga	Ya	2/109=0,0183	0/6 = 0
		Tidak	107/109=0,9816	5/6 = 0,8333
	Depresi	Ya	2/109=0,0183	4/6 = 0,6666
		Tidak	107/109=0,9816	2/6 = 0,3333
	Pasif	Ya	0/109= 0	5/6 = 0,8333
		Tidak	109/109 = 1	1/6 = 0,1666
	Pendiam	Ya	1/109 = 0,0091	4/6 = 0,6666
		Tidak	108/109= 0,9908	2/6 = 0,3333

3) Mengalikan Semua Variabel Perkelas

Dari nilai probabilitas yang telah diperoleh dapat digunakan untuk menentukan label data

testing dengan cara mengkalikan atribut perkelas.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Umur	JK	ode Penya	Waham	Halusinasi	Gaduh Gelisah	Kacau	Sulit Tidur	Bingung	Curiga	Depresi	Pasif	Pendiam	Klasifikasi
68	L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
61	L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
71	L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
62	P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
69	P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	?
62	L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
69	P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
76	P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
71	L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
74	P	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
61	L	F20.3-Und	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	POSITIF
62	L	F20.5-Res	Tidak	Ya	Tidak	Tidak	Ya	Tidak	Tidak	Tidak	Ya	Ya	NEGATIF

Gambar 6. 4 Data Testing

Dari contoh data *testing* di atas kemudian dilakukan perhitungan manual. Hasil perhitungan dapat dilihat sebagai berikut:

Tabel 6. 5 Hasil Perhitungan Manual

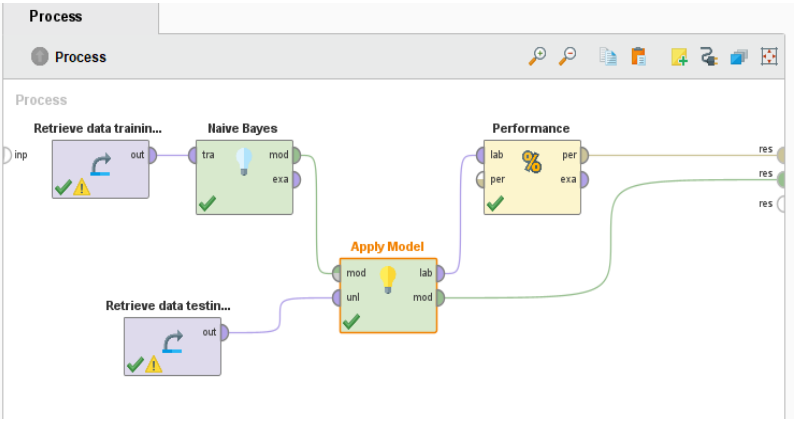
P(X|Ci)	P(JK"=P"|Hasil"Positif")	0,4036
	P(JK"=P"|Hasil"Negatif")	0,5
	P(Diagnosa"=F20.3"|Hasil Positif")	0,9266
	P(Diagnosa"=F20.3"|Hasil Negatif")	0
	P(Waham"=Ya"|Hasil Positif")	0,9724
	P(Waham"=Ya"|Hasil Negatif")	0,1666
	P(Halusinasi"=Ya"|Hasil Positif")	0,9908
	P(Halunasi"=Ya"|Hasil Negatif")	0,3333
	P(Gaduh Gelisah"=Ya"|Hasil Positif")	1
	P(Gaduh Gelisah"=Ya"|Hasil Negatif")	0
	P(Kacau"=Ya"|Hasil Positif")	0,9357
	P(Kacau"=Ya"|Hasil Negatif")	0
	P(Sulit Tidur"=Tidak"|Hasil Positif")	0,9541
	P(Sulit Tidur"=Tidak"|Hasil Negatif")	0,6666
	P(Bingung"=Tidak"|Hasil Positif")	0,9816
	P(Bingung"=Tidak"|Hasil negatif")	0

	P(Curiga"=Tidak"|Hasil Positif")	0,9816
	P(Curiga"=Tidak"|Hasil Negatif")	0,8333
	P(Depresi"=Tidak"|Hasil Positif")	0,9816
	P(Depresi"=Tidak"|Hasil Negatif")	0,3333
	P(Pasif"=Tidak"|Hasil Positif")	1
	P(Pasif"=Tidak"|Hasil Negatif")	0,1666
	P(Pendiam"=Tidak"|Hasil Positif")	0,9908
	P(Pendiam"=Tidak"|Hasil Negatif")	0,3333
P(X|Hasil "Positif")= P(X|Hasil "Negatif")= P(X|Ci)*P(Ci)	0,4036*0,9266*0,9724*0,9908*1*0,9357*0,9541*0,9816*0,9816*0,9816*1*0,9908 0,5*0*0,1666*0,3333*0*0*0,6666*0*0,8333*0,3333*0,1666*0,3333	0,3014 0
P(X|Positif)*P(Positif) P(X|Negatif)*P(Negatif)	0,3014*0,9478 0*0,0521	0,2856 0

Hasil dari perkalian pada *class positif* menunjukan 0,2856. Sedanglan untuk hasil *class negatif* menunjukkan angka 0. Setelah mendapat 2 angka dari *class positif* dan *class negatif* selanjutnya akan mengkomparasi antara hasil dari data uji ini menghasilkan jawaban positif. Selanjutnya adalah menghitung klasifikasi menggunakan algoritma naiva bayes dengan bantuan *software rapid miner*.

b. Pemodelan Menggunakan *Rapid Miner*

Setelah didapatkan hasil pengujian menggunakan *excel*, maka selanjutnya dilakukan pengujian menggunakan *software rapid miner*. Rapidminer merupakan perangkat lunak independen yang digunakan untuk menganalisa data dan mesin penambangan data, yang dapat diintegrasikan dengan berbagai bahasa pemrograman secara mudah (Nasution et al., 2022). Uji coba ini dilakukan untuk mengetahui apakah perhitungan secara manual sesuai dengan perhitungan yang dilakukan dengan *rapid miner*. Berikut merupakan tampilan rapid miner:



Gambar 6. 5 Tampilan Rapid Miner

Dari gambar di atas antara data *training*, data *testing*, naïve bayes, apply model dan performance sudah terhubung satu sama lain. Kemudian klik *icon run* atau segitiga biru pada toolbar untuk menampilkan hasil. Tunggu beberapa saat, komputer memerlukan waktu untuk menyelesaikan perhitungan.

Gambar 6. 6 Icon Run

Setelah beberapa detik maka *rapid miner* akan menghasilkan prediksi pada *view result.* Hasilnya berbentuk tabel seperti gambar di bawah ini:

Row No.	Klasifikasi	prediction(K...	confidence(...	confidence(...	JK	Kode Penya...	Waham
1	POSITIF	POSITIF	1.000	0.000	L	F20.3-Undiffe...	Ya
2	POSITIF	POSITIF	1.000	0.000	L	F20.3-Undiffe...	Ya
3	POSITIF	POSITIF	1.000	0.000	L	F20.3-Undiffe...	Ya
4	POSITIF	POSITIF	1.000	0.000	P	F20.3-Undiffe...	Ya
5	POSITIF	POSITIF	1.000	0.000	P	F20.3-Undiffe...	Ya
6	POSITIF	POSITIF	1.000	0.000	L	F20.3-Undiffe...	Ya
7	POSITIF	POSITIF	1.000	0.000	P	F20.3-Undiffe...	Ya
8	POSITIF	POSITIF	1.000	0.000	P	F20.3-Undiffe...	Ya
9	POSITIF	POSITIF	1.000	0.000	L	F20.3-Undiffe...	Ya
10	POSITIF	POSITIF	1.000	0.000	P	F20.3-Undiffe...	Ya
11	POSITIF	POSITIF	1.000	0.000	L	F20.3-Undiffe...	Ya
12	NEGATIF	NEGATIF	0.000	1.000	L	F20.5-Residu...	Tidak

Gambar 6. 7 Hasil Prediksi Pada View Result

Dari hasil proses perhitungan menggunakan *rapid miner* dengan metode prediksi menampilkan hasil dari data *testing* dan data *training* yang telah diuji. Kolom ini memberikan informasi tentang data pasien skizofrenia pada geriatri yang memiliki gejala positif dan negatif pada tahun 2022. Selanjutnya untuk mengetahui tingkat akurasi algoritma naïve bayes bisa dilihat di *performanceVector* yang terletak di sebalah kanan sehingga hasilnya seperti ini:

accuracy: 100.00%

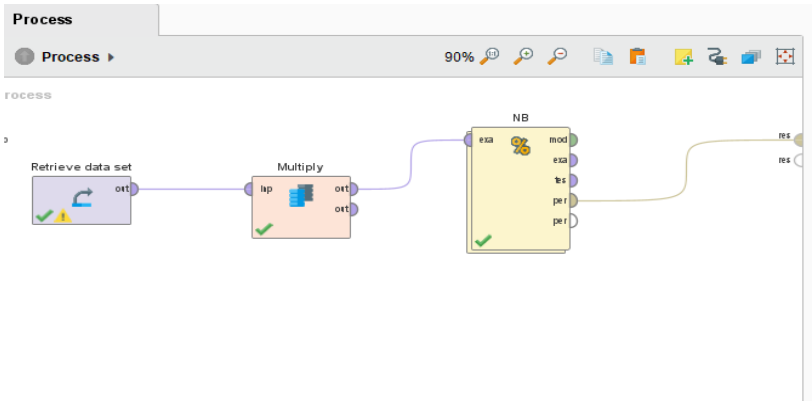
	true POSITIF	true NEGATIF	class precision
pred. POSITIF	11	0	100.00%
pred. NEGATIF	0	1	100.00%
class recall	100.00%	100.00%	

Gambar 6. 8 Performance Vector

Dapat dilihat pada gambar di atas tingkat akurasi dari performanceVector yaitu 100%, class precision yaitu POSITIF 100%, NEGATIF 100% dan untuk class recall yaitu POSITIF 100% dan NEGATIF 100%.

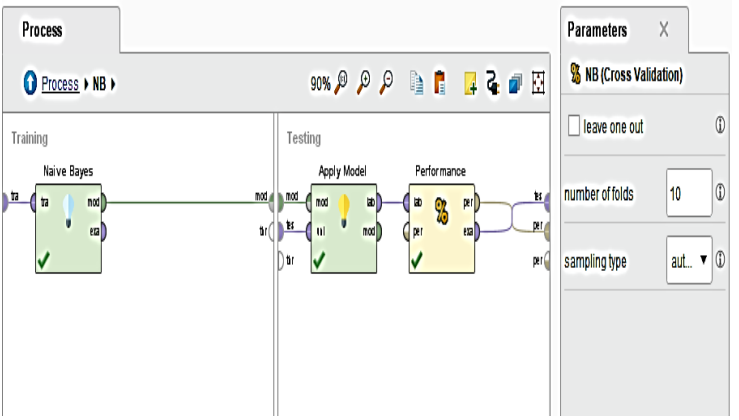
1) Validasi Data (Data *Validation*)

Validasi data dilakukan untuk membantu mengahasilkan tingkat keakurasian berdasarkan dataset yang telah di lakukan proses klasifikasi. Dapat dilihat tampilan eksperimen di R*apidminer*, Dimana *cross validation* digunakan untuk membagi dataset menjadi 10 bagian. Dataset yang dibagi menjadi 10 bagian, terdiri dari 2 jenis yaitu data *training* dan data *testing*. Data *training* fungsinya untuk melatih model sedangkan data *testing* untuk menguji model yang digunakan.



Gambar 6. 9 Tampilan Eksperimen Di Rapidminer

Gambar tersebut menunjukkan eksperimen di *Rapidminer* dimana operator Naïve Bayes merupakan algoritma Naïve Bayes, sedangkan operator *Apply Model* digunakan untuk mengeluarkan *rule* dari *Naïve Bayes* dan operator *Performance* digunakan untuk mengeluarkan hasil kinerja atau akurasi dari model yang digunakan.



Gambar 6. 10 Eksperimen Rapidminer

Hasil kinerja atau hasil akurasi dapat dilihat pada gambar di bawah ini, dimana hasil akurasi dari eksperimen di *Rapidminer* menggunakan algoritma *Naïve Bayes* adalah 98,46.

PerformanceVector (Performance)

Criterion
accuracy

Table View | Plot View

accuracy: 98.46% +/- 3.24% (micro average: 98.43%)

	true POSITIF	true NEGATIF	class precision
pred. POSITIF	118	0	100.00%
pred. NEGATIF	2	7	77.78%
class recall	98.33%	100.00%	

Gambar 6. 11 Hasil Kinerja dari Eksperimen di Rapidminer

Berdasarkan hasil dari serangkaian tahapan penelitian yang telah dilakukan, dapat disimpulkan bahwa proses data mining dengan teknik klasifikasi

menggunakan algoritma naïve bayes dapat diterapkan dalam pengelompokan pasien skizofrenia pada geriatri. Pengelompokan yang dimaksut adalah untuk mengetahui apakah pasien tersebut mengalami gejala positif atau gejala negatif skizofrenia. Dalam perhitungan klasifikasi naïve bayes ini dilakukan secara manual menggunakan microsoft excel dan dengan bantuan software rapid miner. Perhitungan secara manual melalui tahap-tahap yaitu *data selection*, *data preprocessing*, *data transformasi*, *data training* dan *data testing*, kemudian data validasi menggunakan *k-fold cross validation* lalu evaluasi dan hasil.

Dari hasil wawancara dan observasi didapatkan dataset sejumlah 127 data. Data tersebut terdiri dari 120 data pasien bergejala positif dan 7 data pasien bergejala negatif. Setelah itu dari seluruh jumlah dataset dibagi menjadi data training dan data testing. Data *testing* diperoleh dari pembagian sumber data dengan metode 10-*fold cross selection* dimana setiap 10 data latih terdapat 1 data uji. Sehingga pada penelitian ini dengan jumlah dataset sebesar 127, terdapat 115 data latih dan 12 data uji.

Hasil proses klasifikasi pemodelan algoritma naïve bayes terhadap dataset pasien skizofrenia pada geriatri menghasilkan akurasi sebesar 98,46%. Tingkat akurasi yang dihasilkan sudah termasuk kategori sangat baik, namun perlu diadakan peningkatan akurasi dengan perhitungan algoritma lainnya.

DAFTAR PUSTAKA

Ali Muhson. (2018). Teknik Analisis Kualtitatif. *Teknik Analisis Kuantitatif*, 1–7.

Antoni, A., Azizah, G., & Kahayani, D. T. (2016). Tinjauan Kelengkapan Diagnosis Visum Et Repertum Psikiatrik di Rumah Sakit Jiwa Daerah Sambang Lihum Tahun 2015. *Jurnal Kesehatan Indonesia*, *6*(1), 32–38.

Devi Nurul Anisa1, J. (2022). Klasifikasi Penyakit Diabetes Menggunakan Algoritma Naive Bayes. *Dinamika Informatika*, *14*(1), 33–42.

Dzikrina, I. (2021). Penerapan Data Mining untuk Memprediksi Penyekit ISPA. *Informatika*, *6*(1), 1–8.

Heliyanti Susana. (2022). Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet. *Jurnal Riset Sistem Informasi dan Teknologi Informasi (JURSISTEKNI)*, *4*(1), 1–8. https://doi.org/10.52005/jursistekni.v4i1.96

Heriyanto, A. (2021). Penerapan Metode K-Nearest Neighbor (KNN) Untuk Klasifikasi Stanting Pada Balita. *Publikasi Ilmiah Universitas Muhammadiyah Jember*.

Lenaini, I. (2021). Teknik Pengambilan Sampel Purposive Dan Snowball Sampling. *Jurnal Kajian, Penelitian & Pengembangan Pendidikan Sejarah*, *6*(1), 33–39.

Loelianto, I., Thayf, M. S. S., & Angriani, H. (2020). Implementasi Teori Naive Bayes Dalam Klasifikasi Calon Mahasiswa Baru Stmik Kharisma Makassar. *SINTECH (Science and Information Technology) Journal*, *3*(2), 110–117. https://doi.org/10.31598/sintechjournal.v3i2.651

Musu, W., Ibrahim, A., & Heriadi. (2021). Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5. *Prosiding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, *X*(1), 186–195.

Nasution, M., Ritonga, A. A., & Juledi, A. P. (2022). Implementasi Rapidminer dalam Mengklasifikasikan Indeks Demokrasi. *Computer Science and Information Technology (JCoInT)*, *3*(3), 99–106.

Novianti, D. (2019). Implementasi Algoritma Naïve Bayes Pada Data Set Hepatitis Menggunakan Rapid Miner. *Paradigma - Jurnal Komputer dan Informatika*, *21*(1), 49–54. https://doi.org/10.31294/p.v21i1.4979

Pasrun, Y. P. (2019). Penerapan Metode Naïve Bayes untuk Klasifikasi Penyakit ISPA. *semanTIK*, *5*(2), 295–302.

Pebrianti, D. K. (2021). Penyuluhan Kesehatan tentang Faktor Penyebab Kekambuhan Pasien Skizofrenia. *Jurnal Abdimas Kesehatan (JAK)*, *3*(3), 235. https://doi.org/10.36565/jak.v3i3.160

Pradnyana, G. A. (2019). Konsep Dasar Data Mining. *Konsep Data Mining*, *1*, 1–16.

Prastiyo, A. D., & Aji, rojil nogroho bayu. (2020). *Rumah sakit jiwa lawang pascakemerdekaan pelayanan: kesehatan jiwa terhadap pasien ganguan jiwa 1945-1960 dimalang*. *9*(1), 1–10.

Purwono. (2008). Studi Kepustakaan. *Universitas gajah mada*, hal. 66–72.

Rafika Ulfa. (2019). Variabel Dalam Penelitian Pendidikan. *Jurnal Pendidikan dan Keislaman*, *6115*, 196–215. https://doi.org/10.32550/teknodik.v0i0.554

Sudarsono, B. G., Leo, M. I., Santoso, A., & Hendrawan, F. (2021). Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner. *JBASE - Journal of Business and Audit Information Systems*, *4*(1), 13–21. https://doi.org/10.30813/jbase.v4i1.2729

Sugiyono. (2013). *Metode Penelitian Kuantitatif, Kualitatif dan R&D*. Bandung: Alfabeta.CV.

Sukendra, I. K., & Atmaja, I. K. S. (2020). Instrumen Penelitian. In Teddy Fiktprius (Ed.), *Journal Academia*. Mahameru Press.

WHO. (2022). Schizophrenia 10.

Wijaya, A. E., Bani, R., Sukarni, S., & Weighting, S. A. (2019). *Jurnal Teknologi Informasi dan Komunikasi STMIK Subang, Oktober 2019 ISSN: 2252-4517*. (April), 100–110.

Yuli Mardi. (2019). Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Jurnal Edik Informatika. *Jurnal Edik Informatika*, *2*(2), 213–219.

BIODATA PENULIS

Wahyu Wijaya Widiyanto lahir di Kudus, pada 18 September 1986. Ia tercatat sebagai lulusan Universitas Amikom Yogyakarta pada tahun 2019. Wahyu Wijaya Widiyanto bukanlah orang baru di dunia Pendidikan. Ia merupakan lulusan D3 Manajemen Informatika Politeknik Indonusa Surakarta pada tahun 2011, S1 Sistem Informasi STMIK Duta Bangsa Surakarta (2017). Saat ini Wahyu Wijaya Widiyanto aktif sebagai dosen tetap di Politeknik Indonusa Surakarta pada Program Studi Sarjana Terapan Manajemen Informasi Kesehatan. Penulis buku selain bentuk implementasi Tridharma bagian pengajaran sebagai dosen, Wahyu Wijaya Widiyanto juga aktif dalam melakukan penelitian dan pengabdian, hal ini terlihat dari hasil publikasi yang telah dibuatnya, dan dapat dilihat pada akun google scholar di ID NHEh0ZgAAAAJ, ID Sinta 6689713, dan ID Scopus 57215302744. Untuk korespodensi dapat dilakukan melalui e-mail wahyuwijaya@poltekindonusa.ac.id.

BIODATA PENULIS

Bajeng Nurul Widyaningrum, S.Kom, M.Kom yang akrab dipanggil Ajeng lahir di Kendal, pada 3 Juli 1994. Saat ini penulis aktif sebagai dosen prodi Rekam Medis dan Informasi Kesehatan di Politeknik Bina Trada Semarang yang berdomisili di Kota Semarang. Penulis menyelesaikan Studi Sarjana Teknik Informatika pada tahun 2016 dan Magister Teknik Informatika di Universitas Dian Nuswantoro Semarang pada tahun 2019. Tahun 2018 hingga September 2019 pernah bekerja sebagai tenaga kontrak di UPTD Puskesmas Pegandon Kabupaten Kendal. Korespondensi melalui e-mail bnwidyani@gmail.com.